DOCUMENT RESUME

ED 049 608 EM 008 847

AUTHOR	Ferguson, Richard L.
TITLE	Computer Assistance for Individualizing Measurement.
INSTITUTION	Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
PUB DATE	Mar 71
NOTE	90p.; Technical Report
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Branching, *Computer Oriented Programs, Computers, *Criterion Referenced Tests, Elementary School Mathematics, Individual Differences, Individualized Instruction, *Measurement Techniques, Models, *Testing, Test Reliability, Test Validity

ABSTRACT

While the usefulness of branched testing over conventional paper-and-pencil testing has been in doubt, particularly for the student of average ability, this has been with reference to normative measures rather than the criterion-referenced measures characteristic of individualized instruction. A computer-assisted test model for assessing an examinee's proficiency in a set of skills for which a hierarchy of prerequisite relationships is known to exist was developed and evaluated. The test model calls for the random construction of items using item generation rules stored in the computer--an item sampling procedure that permits the test constructor to control for classification errors--and a branching strategy that tailors testing to the individual student in accordance with his competencies. Results with an individualized mathematics program showed the computer test to be highly successful in providing reliable information in substantially less time than was required by conventional methods, even though the sample included students with wide variations in competencies represented by the test unit. The reduction of time required for testing is attributed to the routing strategy rather than the item sampling procedure. (Author/MT)

UNIVERSITY OF PITTSBURGH - LEARNING R & D CENTER

TECHNICAL REPORT	COMPUTER ASSISTANCE FOR INDIVIDUALIZING MEASUREMENT RICHARD L. FERGUSON

EM008947

ED049608

ED049608

COMPUTER ASSISTANCE FOR INDIVIDUALIZING MEASUREMENT

Richard L. Ferguson

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Learning Research and Development Center

University of Pittsburgh

March 1971

This document has been approved for public release and sale, its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the U. S. Government.

The research reported herein was performed pursuant to Contract N00014-67-A-0402-0006 (NR 154-262) with the Personnel and Training Branch, Psychological Sciences Division, Office of Naval Research, with additional funds provided by the National Science Foundation. The document is a publication of the Learning Research and Development Center, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education and Welfare.

TABLE OF CONTENTS

LIST OF TABLES iv

LIST OF FIGURES v

ACKNOWLEDGMENTS vi

ABSTRACT vii

I. THE ROLE OF MEASUREMENT IN THE INDIVIDUALIZATION OF INSTRUCTION 1

1.0 Introduction 1

1.1 Individualized Instruction: The IPI Model 1

1.2 Individualized Measurement: A Rationale for Computer Assisted Testing 3

II. THE NATURE OF INDIVIDUALIZED MEASUREMENT 7

2.0 The General Branched Test Model 7

2.1 Variations in the Branched Test Model 7

2.2 Some Characteristics of the Branched Test Model 12

III. DEVELOPMENT OF A COMPUTER ASSISTED TEST MODEL 17

3.0 Preliminary Specifications for the Test Model 17

3.1 The Decision Component for Proficiency Classifications 20

3.2 The Branching Component 32

3.3 The Item Generation Component 34

3.4 Rationale for Computer Implementation of the Test Model 34

IV. IMPLEMENTATION OF THE TEST MODEL 36

4.0 The Implementation Plan 36
4.1 Development of a Test Hierarchy 36
4.2 Development of Item Generation Rules 39
4.3 The Plan for Proficiency Classification 39
4.4 The Branching Strategy 42
4.5 Administration of the Test 44
4.6 The Collection of Data 47

V. EVALUATION OF THE TEST MODEL 49

5.0 Validity of the Hierarchy 49
5.1 Validity of the Test 50
5.2 Reliability of the Test 51
5.3 Test Length and Time-To-Completion as a Function of Item Sampling 55
5.4 Test Length as a Function of Branching 60
5.5 Some Observations About Implementation of the Computer Assisted Test Model 63

VI. IMPLICATIONS FOR CONTINUING RESEARCH IN MODELS FOR COMPUTER ASSISTED TESTING 67

6.0 Overview 67
6.1 Suggested Refinements for the Test Model 67
6.2 Summary 74

APPENDIX A 78

REFERENCES 80

LIST OF TABLES

Table — Page

3.1 Branching Rules for Computer-Assisted Placement Testing 33

4.1 Objectives for Level D Addition-Subtraction Unit 37

4.2 Sequences for the Level D Addition-Subtracting Hierarchy 43

5.1 Correlation Coefficients Between Repeated Measures of Proficiency for the Seven Linear Sequences 52

5.2 Proportion of Students for Whom Instruction Would Vary from Test to Retest 53

5.3 Proportion of Students Differing from Test to Retest in the Number of Objectives for Which Instruction on a Sequence was Required 53

5.4 Comparison of the Conventional Test with the Computer Test on the Consistency of Proficiency Classification from Pretest to Posttest 54

5.5 Comparison Between Conventional and Computer Tests on the Number of Items Presented 56

5.6 Proportions of Examinees Receiving a Smaller, Equal, or Larger Total Number of Items on the Computer Test than on the Conventional Test 58

5.7 Mean Number of Objectives Tested by Computer for Groups of Varying Proficiency 60

5.8 Number of Objectives Tested Using Two Different Branching Strategies 62

6.1 Average Number of Items Tested Per Objective on the Computer Test 73

LIST OF FIGURES

Figure — Page

1.1 Matrix of Units in the IPI Mathematics Curriculum 4

3.1 Hypothetical Hierarchy for a Set of Five Related Objectives 17

3.2 Graph Illustrating Sequential Probability Ratio Test for Determining Whether a Student Does or Does Not Need Instruction on an Objective . . 30

4.1 Hierarchy of Skills for the Level D Addition-Subtraction Unit 38

4.2 Flow-chart of Item Generation Rules for Objective Eleven 40

4.3 Example of a Student Profile Resulting from the Computer-Assisted Branched Test 45

ACKNOWLEDGMENTS

It would be difficult for the author to adequately express his appreciation to the many persons who contributed to this work. In the case of William Cooley, my major professor, inspirer, and counselor, no expression of gratitude could be sufficient. Similarly, Anthony Nitko was both a thoughtful critic and a helpful advisor throughout the study. Frequent discussions with Robert Glaser provided useful suggestions all along the way. All of this assistance, so generously rendered, helped to improve the quality of the test model that was developed.

Many others have contributed in less academic ways. The administration, staff, and students at the Oakleaf Elementary School of the Baldwin-Whitehall School District were extremely cooperative in providing the support needed to complete the study. My appreciation is also extended to Mrs. Terri Komar, project secretary, whose patience and competence at preparing the manuscript are worthy of special commendation.

This work was made possible by a grant from the Personnel and Training Branch, Psychological Sciences Division, Office of Naval Research. Additional funds were provided by the National Science Foundation. This document is a publication of the Learning Research and Development Center, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education and Welfare.

COMPUTER ASSISTANCE FOR INDIVIDUALIZING MEASUREMENT

Richard L. Ferguson
Learning Research and Development Center
University of Pittsburgh

ABSTRACT

Assuming the existence of a set of skills for which a hierarchy of prerequisite relationships is hypothesized, the computer-assisted test model described in the study is designed to make testing more efficient by capitalizing on the assumed structure.

The model requires that a computer randomly generate items to test a single skill, present those items and score the student's constructed response. The number of items presented before testing of a specific skill is terminated is a function of the examinee's proficiencies and a decision procedure that enables the test builder to specify criteria for sufficient and insufficient proficiency while controlling for Type I and Type II classification errors. A sequential probability ratio test is used to control the latter component of the model.

On the basis of the examinee's cumulative response pattern for a skill and the specification of values for the variables just described, the computer determines whether the examinee has sufficient or insufficient proficiency in the skill or whether an additional item is required before a decision can be made. If a proficiency decision can be made, the examinee is routed for testing on another objective. Routing is based on criteria specified by the test constructor in accordance with his knowledge of the hierarchy for the skills. Otherwise, another item is generated and the cycle repeated.

Results showed that the computer test was highly successful in providing reliable information in substantially less time than that which was required by the conventional paper-and-pencil test. Reduction of the time required for testing is attributed to the routing strategy rather than the item sampling procedure. The test model proved to be extremely effective for all examinees involved in the study, even though the sample included students with wide variations in the competencies represented by the test unit.

COMPUTER ASSISTANCE FOR INDIVIDUALIZING MEASUREMENT

Richard L. Ferguson

Learning Research and Development Center

University of Pittsburgh

I. The Role of Measurement in the Individualization of Instruction

1.0 Introduction

Educational systems such as Individually Prescribed Instruction (Glaser, 1968) and A Program for Learning in Accordance with Needs (Flanagan, 1967) have demonstrated the feasibility of instructional programs that are designed to be adaptive to the unique requirements of individual learners. These programs accomplish individualization in a variety of ways. Permitting students some choice in determing the skills they will learn, developing alternative instructional sequences for teaching skills, and establishing organizational procedures that permit students to progress at different rates, are examples of how such programs achieve truly individualized educational environments. Since the subject of this report is closely allied to individualized learning systems, the Individually Prescribed Instruction (IPI) model is described in some detail in succeeding sections of this paper.

1.1 Individualized Instruction: The IPI Model

Individually Prescribed Instruction (IPI) is a project of the Learning Research and Development Center (LRDC) at the University of

Pittsburgh. Cooley and Glaser (1969) define individualized education as "the adaptation of instructional practices to individual requirements." Three major components are involved in their conceptualization of an individualized system: (1) educational goals, (2) individual capabilities, and (3) instructional means. Their model of instruction attends to these components by way of a sequence of operations that more precisely defines the IPI model (Glaser, 1968). They include:

(1) Specification of the skills to be learned in terms of observable student behavior.

(2) Assessment of an individual's skills upon entry to a course of instruction.

(3) Assignment or election of educational alternatives fitted to the student's entering proficiencies.

(4) Continuous assessment and monitoring of the student's performance and progress.

(5) Completion of available instructional sequences as a function of assessment of student performance and criteria for proficiency.

(6) Collection of data for improving the instructional system.

The successful implementation of an instructional model such as IPI requires that a teacher manage the learning activities of a large number of students, many of whom have widely varying competencies and needs. Ferguson (1970) observes that although IPI solves the immediate problem of providing instructional sequences for students with diverse needs and skills, it does not obviate the problem of managing instruction. He notes that "it shifts the emphasis of

management from providing a lesson for a student to providing the optimum instructional strategy for him given the maximum amount of information that would be useful in making such a choice and the restrictions imposed by available instructional resources."

Glaser (1968) notes that in the IPI educational model, test data serves as the primary source of information enabling teachers to make differential decisions regarding instruction for individual children. Prior to that observation, Glaser (1963) pointed out the need for designing and using criterion-referenced measures to assess learning outcomes. Such measures are used to estimate the proficiency that an individual has attained in a skill in terms of some specified criteria. Criterion referenced tests or performance measures have been used with considerable success in IPI. However, it seems clear that as the IPI model of instruction undergoes continuous refinement, changes in testing procedures will also follow.

1.2 Individualized Measurement: A Rationale for Computer-Assisted Testing

Bolvin (1967) provides a description of both the IPI mathematics curriculum and the testing procedures used to diagnose student difficulties and monitor their progress. A brief description of these two components of the IPI mathematics program may be useful.

Figure 1.1 conveys the general organization of the mathematics curriculum. Thirteen content areas, numeration, place value, addition, etc., are identified; each occurring at various levels of complexity. The intersection of each level with a specific content area determines a unit that consists of a set of behaviorally defined

objectives or skills. Thus, level A numeration is a unit that consists of a set of objectives that share a similar content but are less complex than the objectives contained in the level B numeration unit. The absence of an X at any position in the chart indicates that no unit exists for the corresponding content area and level.

Figure 1.1

Matrix of Units in the IPI Mathematics Curriculum

	LEVEL							
	A	B	C	D	E	F	G	h
Numeration	X	X	X	X	X	X	X	X
Place Value		X	X	X	X	X	X	X
Addition	X	X	X	X	X	X	X	X
Subtraction	X	X	X	X	X	X	X	X
Multiplication				X	X	X	X	X
Division				X	X	X	X	X
Combination of Processes			X	X	X	X	X	X
Fractions	X	X	X	X	X	X	X	X
Money		X	X	X	X	X		
Time	X	X	X	X	X	X	X	
Systems of Measurement		X	X	X	X	X	X	
Geometry		X	X	X	X	X	X	X
Special Topics			X	X	X	X	X	X

Prior to undertaking work on a unit, a child is administered a pretest that provides a measure of his proficiency in each objective in the unit. Since testing is criterion-referenced, the primary output from the pretest is a list that classifies the examinee as to his competency in each of the objectives. The latter decisions are obtained in accordance with a pre-determined rule that is independent of the performance of other students. After instructional sequences

are completed for all objectives in which the examinee does not have sufficient proficiency, he is administered a unit posttest, an equivalent form of the pretest. If proficiency criteria for all of the objectives are met, he proceeds to another unit; otherwise, he is usually given additional instruction and eventually another posttest.

The implementation of individualized learning environments has typically produced large variation in the rates at which students progress through a curriculum. Glaser (1968) found substantial variance in the number of units completed by elementary school students who had participated in three years of individualized instruction in mathematics. Similarly, Suppes (1964) reported a wide range in the rate of learning for first grade students studying mathematics. Since the entering competencies that students possess vary when they begin study in a specific mathematics unit, not all students would benefit from the same instruction. Consequently, the content and quantity of instruction are adjusted to the individual need, thus contributing to the variance in rates for completion of a unit. It seems natural to assume that, just as instruction should be adaptive for students possessing different competencies, measurement should also be individualized. For example, in order to identify the individual competencies of a group of students in a given unit, it should not be necessary to test all students on all skills.

A close examination of the units in the mathematics curriculum reveals the existence of prerequisite relationships for the objectives within and among units. Such an observation led to the hypothesis

that some form of sequential testing might prove useful for IPI mathematics. A procedure could be developed that, with the aid of a computer, would tailor measurement of an individual so as to include only those objectives or competencies that were neither trivial nor unsolvable for him. Branched tests, tailored tests, and sequential-item tests are names that have been used to describe instruments having this attribute.

The purpose of the present study was to develop a model for computer assisted branched testing and to investigate the problems associated with its implementation in a program of individualized mathematics instruction. Therefore, its primary focus was upon the development of measurement procedures that would accommodate a more effective instructional program.

II. The Nature of Individualized Measurement

2.0 The General Branched Test Model

In the interest of establishing a common understanding for the term branched test, the latter is defined to be any instrument designed to measure a set of skills or objectives by routing the examinee to items neither too easy nor too difficult for him to solve. Thus, in IPI mathematics, where students are tested on all skills comprising a unit, a branched pretest or posttest would limit testing to include only those skills of appropriate difficulty for the individual student.

2.1 Variations in the Branched Test Model

Differences among the branched test procedures that have been reported to date are defined by variations in the manner in which routing is accomplished or in the number of items presented to measure a specific objective prior to routing. A description of seven branched test procedures developed by Cleary, Linn, and Rock (1968) will serve to illustrate these differences. Their tests were developed using existing item-response data from 190 verbal-type items for nearly 5000 examinees. Since very examinee had responded to all 190 items, it was possible to simulate each of the branched test procedures using the available data.

Five of the seven procedures were used to develop tests with two distinct components: (1) a routing section, and (2) a measurement section. On the basis of an examinee's responses to items that

comprised the routing section of each of the tests, he was branched to one of four measurement sections of varying difficulty. Each of the five tests employed a different routing procedure; that is, used a different strategy to determine the measurement section on which an examinee would be tested. All five routing methods used student responses to approximately 20 of the 190 items as a basis for assigning students to a measurement section of the test. Each of the measurement sections was comprised of the 20 items with the highest within-group point-biserial correlations, excluding any of the 190 items used for the routing tests. Thus, regardless of the particular routing method used, each examinee was assumed to have been branched to one of four 20-item measurement sections as a function of his prior responses to a common 20-item routing section of the test.

The two remaining test procedures did not follow the two-stage design of the other five. Whereas routing occurred only once during testing with each of the previous procedures, in the latter two cases it occurred after each item or block of items had been presented. The first method called for an examinee to be branched to another item as a function of the correctness of his response to the last item that he was presented. A correct response resulted in a branch to a more difficult item; an incorrect response to a less difficult item. With the second procedure, branching occurred after a block of five items, each of comparable difficulty, was completed by the examinee. Thus, both multiple routing procedures called for the outcome of measurement at a specific level of

difficulty to be used as input for making a decision regarding what item or block of items should be tested next.

Each of the seven test procedures was simulated using the existing data and the results evaluated. The five two-stage branched tests were found to be quite successful when reproducibility of the 190-item total test score was the criterion. However, to facilitate comparison of the branched test method with conventional test procedures, shortened conventional tests were also scored. The latter were constructed using items from among the 190 with the highest point-biserial correlations with the total test score. With respect to the reproducibility criterion, the two-stage branched tests were found to be only slightly superior to shortened conventional tests of comparable lengths.

When the multiple-routing testing procedures were evaluated in a similar fashion, the procedure that called for branching to occur after each item proved to be slightly less adequate in terms of the total score reproducibility criterion than a conventional test of the same length. However, the procedure by which branching occurred after every five items proved to be better able to meet the reproducibility criterion than a conventional test of twice its length.

If one views the items required to test a single IPI objective as a block, then branching from one block of items to another is isomorphic to branching from one objective to another. Building on this assumption, it is clear that a multiple routing procedure

that requires responses to a block of items prior to branching, appears to offer an efficient alternative to current IPI mathematics pretesting and posttesting procedures. However, several adjustments in the strategy should be investigated. One is tempted to inquire into the desirability of estimating an individual's proficiency in a single skill by using a block of n items. For some examinees, it is probably the case that an accurate decision could be made with fewer than n items, whereas for others, n items may be inadequate.

The formulation of a viable branching strategy, one which is adaptive to the cumulative performance of an examinee, is another concern that requires attention. In the study just described, one procedure called for branching an examinee to a block of items of lesser or greater difficulty as a function of the number of correct responses he had made on the last block of items. Such a strategy, although superior to the current pretest procedure that requires that all objectives comprising a unit be tested, could be improved even further if the branching plan incorporated a knowledge of the prerequisite relationships among the unit objectives.

In the IPI context, if a hierarchy of prerequisite relationships was established for a set of skills, branching rules could be defined such that, if the examinee lacked sufficient proficiency in the last objective tested, the routing strategy would direct him to a test on a prerequisite objective. If he had sufficient proficiency in the skill, he would be branched and tested on a higher

order objective. Branching and testing would continue until the examinee's status in all objectives was established. In this way, the actual amount of testing required would be reduced substantially. Since the role of testing in IPI mathematics is to provide information for making instructional decisions, a large amount of each student's time is necessarily allotted to testing. A testing procedure that would, at a minimum, provide the same information as the conventional pretests and posttests, but in much less time, would contribute substantially to improvement of the instructional model.

The methods studied by Cleary et al., although somewhat artificial because they were simulations that used data generated prior to the study, encourage the investigation of similar testing methods for IPI mathematics. The multiple routing method that they investigated is of particular interest as it proved to be quite successful in terms of reproducing total test score while reducing the number of test items.

In a somewhat different approach, Kriewall and Hirsch (1969) report the use of an item sampling model for testing three simple skills involving operations with fractions. Items for the three tests were developed using item-generation rules. Five items were selected from the item population for each test. Nineteen fifth-grade children were administered five-item pretests and posttests for each of the three skills. Having fixed the number of items at five, and the error criterion at two, they determined that to

discriminate a maximum error rate of .15 and a minimum error rate of .65, the errors of classification would be α=.16 and β=.04. As would be expected, errors of classification would decrease as the number of items in each test increased. An item sampling procedure that would permit some control over incorrect proficiency classifications could yield substantial payoffs, particularly, when tests are used to make decisions regarding instruction.

2.2 Some Characteristics of the Branched Test Model

Numerous researchers (Bayroff and Seeley, 1967; Angloff and Huddleston, 1958; Hanson and Schwarz, 1968; Waters, 1964; and Patterson, 1967) have undertaken the study of test characteristics for measurement instruments that incorporate routing procedures. Generally, they concluded that the branched tests with which they experimented were well endowed with reliability, although their indices were derived in a variety of ways. A review of their combined efforts leads to a cautious optimism as to the potential for branched testing in a program of individualized instruction. Generally, the studies reported were not conceived with an interest in examining the instructional implications for branched testing procedures. To the contrary, most were focused upon determining whether or not such instruments could adequately reproduce conventional achievement test scores. As a consequence, much remains to be learned with respect to their potential in an instructional system.

In a recent discussion, Lord (1968) was concerned with some theoretical questions that are encountered when dealing with branched testing. The focus of his study, like those previously discussed, was on measurement as an end in itself, rather than on measurement having instructional implications. The specific problem to which his discussion was addressed was the measurement of a single individual with respect to one psychological dimension. Of basic concern was the selection of n test items that, having been administered to the examinee, provide n responses that were used to estimate the psychological dimension to be measured.

One major conclusion that Lord reached was that in a typical test, where item difficulty is not too heterogeneous, the majority of items are already well tailored to the abilities of most of the examinees. This being the case, he concluded that tailored testing offers little hope for improving measurement for most examinees.

A study by Waters (1964) lends additional support to Lord's observation. She concluded that, if the number of items presented on a conventional test and a branched test began with some small fixed number and increased, the latter would prove superior for measurement purposes initially but this superiority would be maintained only until the conventional test provided a sufficient number of items near the examinee's ability levels.

Lord's discussion and the others that were reported must be considered in terms of their limitations, lest they cast an unwarranted sense of futility on attempts to benefit from branched

testing procedures. It must be re-emphasized that they were not concerned with designing instruments that could provide input for making decisions effecting instruction. Rather, Lord's discussion assumes a relatively homogeneous set of items, quite unlike what would be encountered in an IPI unit. As a consequence, his discussion does not attend to what may well prove to be a very significant application of branched testing, criterion referenced measurement directed at providing input for the formulation of instructional strategy.

Green (1968) lends further credence to this hypothesis. He observes that computer-based branched testing may prove to be a valuable tool in permitting interplay between instruction and evaluation. He states:

> Lord's gloomy conclusions about tailored testing arise from considering measurement per se, rather than any use to which the measures are put. This restricted outlook is in tune with our current wasteful decision-making procedures in education, industry, and the military establishment. We typically measure first and decide later. There is seldom any interplay between measurement and decision. No allowance is made for a decision to collect more data.

He further notes that when a test is to be used for placing students on a hypothetical ability dimension, such as would be required in IPI mathematics, a branched test would be preferred to a conventional test since each examinee would be asked to answer only those items that provide information on individuals at his level of proficiency. In addition, he notes that it is inevitable that branched tests, which have items of varying difficulty, be inferior for measurement

of the student of middle ability since wasted movement is certain when attempting to determine the difficulty level upon which to focus. Further, he emphasizes that an advantage of tailored testing is its ability to stop the testing process as soon as a decision can be made, thus placing the focus upon tailoring the number of items, not their difficulty.

This latter observation is not entirely appropriate if one is considering branched testing as a procedure for within unit placement testing in IPI mathematics. In this situation, it is highly desirable that the branched test permit tailoring of both the number of items presented and the level of their difficulty. The number of items required to determine the proficiency status of an examinee for a particular skill would be varied in accordance with his unique competencies at the time of testing. Once testing was completed for a single skill, the difficulty of the next skill to be tested would be tailored to the individual as a function of his performance on the skills previously tested.

It should be added that although both of these adaptive functions could be carried out in a non-automated manner, it seems highly probable that only a computer could make the interaction required for such a test procedure practical and also make its wide-scale implementation feasible.

In summary, branched testing appears to have a potentially useful application in providing the kinds of information essential for making instructional decisions. Although some evidence has

accrued to justify this optimism, no substantive study has attempted to relate this testing procedure to an instructional program.

On the occasions when branched tests have been administered, the usual focus has been upon tests for which the goal is normative measurement. Such studies have made it clear that for most examinees, branched testing offers no substantive improvement over conventional tests. Previous studies indicate that branched testing appears useful in a measurement sense only when the examinee's ability lies somewhere near the extremes of the continuum. The object of this study then, is to estimate the impact that computer-assisted branched testing might have on a program of instruction that relies heavily upon test data for determining instructional strategy; for clearly, to improve testing without improving the instructional process would be at best a hollow victory.

III. Development of a Computer Assisted Test Model

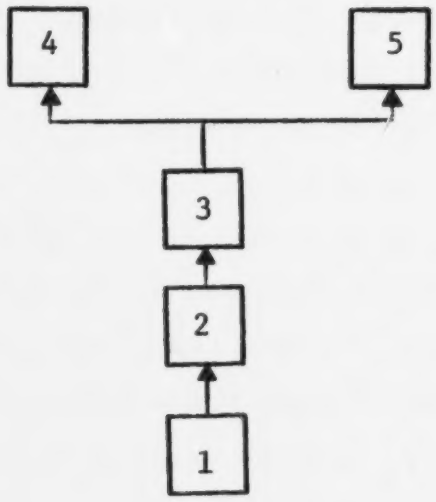
3.0 Preliminary Specifications for the Test Model

Practical demonstrations of the capacity of computers to contribute to the solution of measurement problems are almost non-existent. Consequently, it is difficult to do anything other than speculate as to the potential of computers for performing such roles. This study proposed to develop and implement a computer assisted test model that can be applied to any situation which requires that a set of 'related' skills be tested. The term related is used in the sense that prerequisite relationships are known to exist among the skills.

To clarify the latter notion, Figure 3.1 offers a graphic representation of the prerequisite relationships among a hypothetical set of objectives to which the test model might be applied.

Figure 3.1

Hypothetical Hierarchy for a Set of Five Related Objectives



For the five objective hierarchy, objective 1 is prerequisite to all other objectives; that is, proficiency in objective 1 is necessary but not sufficient for proficiency in the remaining objectives. Whereas, objective 2 is prerequisite to objectives 3, 4, 5, and objective 3 is prerequisite to objectives 4 and 5, the latter two objectives represent terminal behaviors for the unit and neither is prerequisite to the other.

Development of the test model was guided by the desire that it possess several characteristics, many of which could not easily be implemented without the assistance of a computer. These test attributes are categorized and specified below.

Classification Decisions

(1) The test user can vary the criterion levels used to determine an examinee's state of proficiency in a particular objective. Example: As testing proceeds, item responses are processed on a one-by-one basis. An examinee is said to evidence sufficient proficiency in an objective if the estimate of his true proficiency in the objective, based on his responses to the items processed to the given point in testing, is greater than or equal to some value p_0, say .90. Likewise, a minimum value p_1 could be set such that if the estimate of the examinee's true proficiency in the objective is less than, say .50, he is said to have insufficient proficiency. The values for p_0 and p_1 must be subject to manipulation by the test constructor; that is, he must have the capability to vary them among objectives.

(2) When measuring an examinee's proficiency in a particular objective, the test user can specify and impose limits for tolerable risks of misclassifying the examinee. Example: Testing of an examinee might continue on objective 4 until a decision on his proficiency was made that risked Type I and Type II errors no greater than .10 and .05, respectively. As in (1), the risk of classification errors should be free to vary among objectives.

Branching Strategy

The examinee is not tested on the entire set of objectives. Rather, he is branched from objective to objective in accordance with the existing prerequisite structure for the set of objectives at hand and the cumulative record of his performance on each objective measured to the given point in testing. Example: Testing for an examinee might begin with objective 3. Should the examinee be classified as having insufficient proficiency in the objective, he would not be tested on any objectives to which 3 was prerequisite, in this case, objectives 4 or 5. Rather, he would be branched for testing on either objective 1 or 2.

Item Generation

Since it is assumed that the same number of items are not required to determine proficiency in a particular objective for every examinee, and since, for instructional settings, unique but equivalent forms of a test are often desirable, items will be constructed at the time they are needed by using item generation rules stored in the computer.

Succeeding sections of this paper elaborate on the test attributes just described.

3.1 The Decision Component for Proficiency Classifications

If test information is to be used to formulate instructional strategy for individual students, the procedures by which proficiency decisions for specific objectives are made should provide for two contingencies. First, the test user, be he teacher or curriculum expert, should be able to specify precise criteria for making decisions regarding an examinee's relative proficiency in a given objective. Second, he should be able to control the probability of his committing an error in classifying the examinee with respect to the latter's proficiency in the objective.

It is often the case that the size of the population of items required to exhaustively test an objective is very large. For many IPI mathematics objectives, the size of the population is numbered in the hundreds or thousands. Obviously, the latter is a function of the precision with which curriculum developers specify the objectives. For example, consider the following objectives:

(1) Subtracts single digit addends that yield a single digit sum.

(2) Subtracts single digit addends.

The first objective can be exhaustively tested with 55 items, the number of items that results when each of the ten digits is subtracted from every other digit (excluding negative results) and itself. Since the item population for the first objective is a subset of the

item population for the second, exhaustive measurement of the former could be accomplished with fewer test items that would be required for the latter. The decision as to the precision with which an objective is stated is a matter of instructional consequence, and accordingly should be made in terms of instructional considerations. If dividing the second objective stated above into two separate objectives, one of which is equivalent to the first objective, would facilitate learning of the skills represented, then such an action should be taken.

The task for the test constructor is to develop a test that measures the individual's proficiency in a stated objective, regardless of its form. Since it is usually neither efficient nor practical to test the entire population of items for a particular objective, an estimate of an individual's proficiency in an objective can be obtained by employing an item-sampling process that tests items that have been randomly selected from the population of items defining that objective.

Setting the proficiency criteria for an objective is viewed as a somewhat arbitrary action. Proponents of criterion referenced measurement vary in their view as to how rigid these standards should be. Arguments are heard that favor classifying an examinee as proficient in a skill only if he responds correctly to all items on a test that measures the skill. Others seem ready to accept less demonstrative test performances as indicative of sufficient proficiency in a skill.

With the test model under consideration, the test user is free to specify a minimum performance criterion, say p_0, that must be met by the examinee if it is to be said that he has sufficient proficiency in the skill. It is further possible to permit him to vary that criterion from skill to skill in the same test. In addition, he can specify a second performance criterion, say p_1, with $p_1 < p_0$, such that if it is determined at any point in testing that the examinee has failed to achieve a level of performance greater than or equal to p_1, he is said to have insufficient proficiency in the skill and testing is terminated. If application of the two criteria fail to yield a decision in either direction, an additional item testing the same skill is administered and the classification criteria re-applied.

To summarize, assume that for purposes of discussion, 'm' items from the population of items for some objective has been randomly generated, administered to an examinee, and his responses processed. Prior to the test, curriculum authorities agree that the criteria to be used to reach a proficiency decision are $p_0 = .85$ and $p_1 = .60$. Therefore, on the basis of the examinee's responses, it can be stated that if $p \geq p_0$, where p is the percentage of items that the examinee has answered correctly to the given point in testing, he should be classified as sufficiently proficient in the skill. However, if $p \leq p_1$, the decision is that, on the basis of his performance to this point in testing, he lacks sufficient proficiency in the skill. If $.60 < p < .85$, judgment is reserved until an

additional item is generated and the procedures for classification are repeated.

It should be clear that the values chosen for p_0 and p_1 are arbitrary. If the skill that is being tested is extremely important to a student's future progress in the curriculum, then values for p_0 and p_1 might be set at .95 and .80, respectively. Other circumstances might suggest the use of less rigid criteria.

It has been suggested that the number of items required to test different examinees on the same objective should be variable. This is intuitively appealing since one could not expect all examinees to possess identical proficiency in any particular skill. Fewer items should be required to classify the examinee who has no competency in a skill than to classify the examinee who is extremely competent in the skill.

The issue that must be addressed is, how many items must be sampled in order that a particular proficiency decision can be made with some specified confidence that the decision will not result in a mis-classification of the examinee? Thus, two additional variables enter the discussion. They are, the probabilities of Type I and Type II classification errors.

Any sampling plan that does not exhaust the population of items testing a given objective, may lead to an incorrect decision regarding an examinee's proficiency in the objective. Since exhaustive testing is impossible, it is necessary to function with the risk of making decisions that result in mis-classification. In

defining an item sampling plan, it will be necessary to specify the maximum risks of incorrect classification decisions that can be tolerated. In an instructional context, a Type I (α) error occurs when an examinee is sufficiently proficient in a skill but test results yield an opposing classification. As a result, he is prescribed work lessons that may serve no useful function. A Type II (β) error occurs whenever an examinee in fact lacks proficiency in an objective, but on the basis of test results is classified as having sufficient proficiency. The consequence of this error is that needed instruction is not provided. In IPI mathematics, a Type II error is perceived to be potentially more serious than a Type I error since the former could easily result in a child having difficulty proceeding through a unit and might eventually lead to an impasse in instruction; whereas, the latter will at worst require that the student pursue a review-like study of skills in which he is already proficient.

An item sampling plan that addresses this concern for building statistical confidence into the decision process used to classify an examinee as to his proficiency in an objective can be described by a Bernoulli-type experiment, the results of which can be fitted to a binomial distribution. The assumptions of such an experiment are three in number: (1) The possible number of outcomes for each trial must be precisely two, (2) the probability of each of the outcomes must be constant over trials, and (3) the outcome of any trial must be independent of the outcome of all other trials.

Since one may perceive of the response to an item as either correct or incorrect, the first assumption holds. Further, since the items are intended to measure a specific behavioral objective, it is assumed that each item contributes the same information to the classification decision as all other items. Finally, it can be observed that the response to any item is not dependent upon the response to previous items for the same objective; that is, the items enjoy local independence.

The proposed test model assumes that at any given moment in time, a single numerical value can be used to represent the proficiency of an examinee with respect to a specified objective. His relative true score on the population of items is an estimate of this proficiency. Initially, an examinee is presented with an item that is randomly generated from among the population of items for the objective being tested. After he responds to the first item, his response is scored as either correct or incorrect. At this point, a decision is made that, with some arbitrarily-fixed risk of error, classifies the examinee as either having or not having sufficient proficiency in the skill. A third possibility is that the classification decision is deferred until another item has been presented and the item response processed. The latter action is required when an insufficient number of items have been presented to make a decision that satisfies the specified error criteria.

A sampling plan that satisfies the conditions thus far specified is given by the sequential probability ratio test (Wald, 1947 of strength (α,β) for testing the hypotheses:

1) H_0: $p = p_0$

2) H_1: $p = p_1$

In the model, p is the unknown proportion of items that would be answered correctly if testing were over the entire population of items for the objective. p_0 and p_1 are classification criteria that are chosen arbitrarily, where p_0 is larger than p and p_1 is smaller than p. p_0 and p_1 are specified such that indicating sufficient proficiency in the objective is an error of grave consequence only if $p \leq p_1$ and indicating insufficient proficiency in the objective is a serious error only if $p \geq p_0$. If p is situated between p_0 and p_1, the decision is deferred until another item is processed and thus no error can occur.

The risks that are taken are specified in the following manner. The probability of declaring insufficient proficiency in the objective should not exceed some small predetermined value α whenever $p \geq p_0$ and the probability of declaring sufficient proficiency in the objective should not exceed some small value β whenever $p \leq p_1$. It becomes clear that control of the decision process resides in large part with the test user since he is free to vary p_0, p_1, α, and β, to suit a given instructional situation.

For the purpose of discussion, suppose that the following values are assigned to the preceding variables by the test user. Let $p_0 = .85$, $p_1 = .60$, $\alpha = .20$, and $\beta = .10$. Then, one can state that if a proficiency decision is reached after an examinee's responses to $\underline{m}$ items have been processed, the probability that

he will be incorrectly classified as having sufficient proficiency in the skill will not exceed .20 whenever $p \geq .85$ and the probability that he will be incorrectly classified as having sufficient proficiency in the skill will not exceed .10 whenever $p \leq .60$.

The test for classifying an examinee as to his level of proficiency in an objective is described as follows:

Let x_i represent the evaluation of the response to the $i^{\underline{th}}$ item where $x_i \in U$ and:

$$U = \begin{cases} 1 \text{ if the } i^{\underline{th}} \text{ item is answered correctly} \\ 0 \text{ if the } i^{\underline{th}} \text{ item is answered incorrectly} \end{cases}$$

Assume that testing for a particular objective has progressed to the stage that $\underline{m}$ items have been presented and the examinee's responses processed. The probability that the obtained responses for the $\underline{m}$ items would yield a sample equal to $(x_1, x_2, \ldots, x_m)$ is $p^{c_m}(1-p)^{w_m}$, where $c_m = \sum_{i=1}^{m} x_i$; that is, the number of items in the sample of size $\underline{m}$ answered correctly, and $w_m = m - c_m$.

Under H_0: $p = p_0$, the probability of the same sample becomes $p_{0_m} = p_0^{c_m}(1-p_0)^{w_m}$ and under H_1: $p = p_1$, the probability becomes $p_{1_m} = p_1^{c_m}(1-p_1)^{w_m}$. The sequential probability ratio test is then applied in the following manner. At each stage in testing, after the examinee has responded to the $\underline{m}^{th}$ item and his response scored as a zero if incorrect and a one if correct, the following computations are carried out:

$$\log \frac{p_{1_m}}{p_{0_m}} = c_m \cdot \log \frac{p_1}{p_0} + w_m \cdot \log \frac{1-p_1}{1-p_0}$$

where $\underline{m}$ is the number of items tested thus far. Testing for an objective continues as long as:

$$(1) \quad \log \frac{\beta}{1-\alpha} < \log \frac{p_{1_m}}{p_{0_m}} < \log \frac{1-\beta}{\alpha} .$$

Testing ceases as soon as the preceding inequality fails to hold. If at that point:

$$(2) \quad \log \frac{p_{1_m}}{p_{0_m}} \geq \log \frac{1-\beta}{\alpha} ,$$

the examinee is said to have insufficient proficiency in the objective. If instead,

$$(3) \quad \log \frac{p_{1_m}}{p_{0_m}} \leq \log \frac{\beta}{1-\alpha} ,$$

the examinee is said to have sufficient proficiency in the objective.

The preceding three inequalities are equivalent to the following:

$$(1') \quad \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{1-p_1}{1-p_0} - \log \frac{p_1}{p_0}} + m \cdot \frac{\log \frac{p_0}{p_1}}{\log \frac{1-p_1}{1-p_0} - \log \frac{p_1}{p_0}} < w_m <$$

$$\frac{\log \frac{1-\beta}{\alpha}}{\log \frac{1-p_1}{1-p_0} - \log \frac{p_1}{p_0}} + m \cdot \frac{\log \frac{p_0}{p_1}}{\log \frac{1-p_1}{1-p_0} - \log \frac{p_1}{p_0}}$$

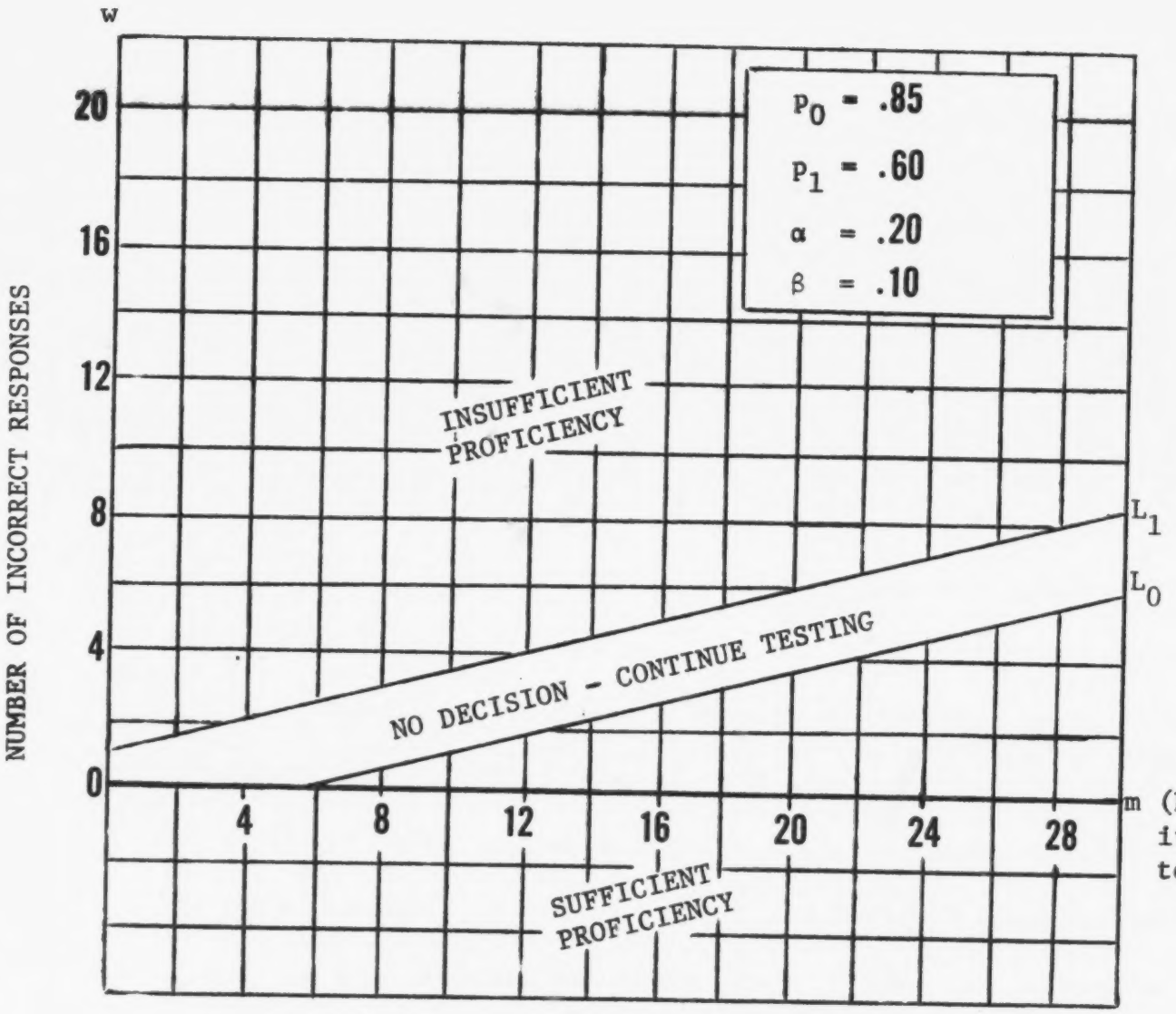
(2') $w_m \leq a_m$, where a_m is the right member of the inequality in (1').

(3') $a_m \geq r_m$, where r_m is the left member of the inequality in (1').

a_m will be referred to as the acceptance number (retain H_0) and r_m as the rejection number (reject H_0). Thus the procedure can be described in the following manner. After each test item is scored, a_m and r_m are computed. Testing continues if $a_m < w_m < r_m$. If $w_m \geq r_m$, insufficient proficiency is indicated and if $w_m \leq a_m$, sufficient proficiency is indicated. Thus, the acceptance number a_m and the rejection number r_m are clearly dependent upon p_0, p_1, α, and β.

Since the inequalities associated with a_m and r_m can be thought of as occupying two mutually exclusive areas in a Cartesian space, it is possible to present a graphic representation of the decision process resulting from an application of the sequential probability ratio test. Figure 3.2 provides such a chart for an arbitrary set of values for p_0, p_1, α, and β. When graphed, the equations $w_m = a_m$ and $w_m = r_m$, yield the lines L_0 and L_1, respectively. All points on the graph below L_0 determine an area of sufficient proficiency. Thus, an examinee who has responded correctly to 8 of 10 items, that is, for whom $m = 10$ and $w = 2$, is said to have sufficient proficiency in the objective. Similarly, all points above line L_1 on the graph represent an area of insufficient proficiency.

Figure 3.2 facilitates the description of how item sampling proceeds. Recall that in the test model, p denotes the unknown



H_0: p=.85 (Student has sufficient proficiency, omit instruction)

H_1: p=.60 (Student does not have sufficient proficiency, give instruction)

Figure 3.2

Graph Illustrating Sequential Probability Ratio Test for Determining Whether a Student Does or Does Not Need Instruction on an Objective (Modified from Ferguson, 1969)

proportion of items that would be answered correctly if testing were over the entire population of items for the objective. p_0 and p_1, which may be varied for different objectives, are selected by curriculum specialists. Thus, if $p_0 = .85$ and H_0 (p=.85) is retained, the examinee is said to have sufficient proficiency in the objective and no instruction is prescribed. However, if $p_1 = .60$ and H_1 (p=.60) is retained, insufficient proficiency is declared and a need for instruction is indicated.

As can be seen on the graph, neither a correct nor incorrect response on the first item could result in a decision that would end testing of the objective. If m, the number of items tested to date, equals one, then regardless of the value of w, the decision for the objective is that testing should continue. A minimum of two items are required before a decision is possible. If the examinee answers the first two items incorrectly, H_0 is rejected, H_1 is retained, and he is said to have insufficient proficiency in the objective. The shortest route leading to a decision for sufficient proficiency with $p_0 = .85$ is a correct response to each of the first six items. In general, testing continues with the generation and presentation of another item whenever $.60 < p < .85$. Since it is possible for an examinee's true proficiency to lie in the "no decision" region, a rule is applied to truncate the testing cycle after some arbitrary maximum number of items has been tested. In the event of truncation, the classification rule holds that the examinee has sufficient proficiency in the objective if after the last item,

the decision function fixes the point (m,w) on the graph closer to the sufficient proficiency region than to the insufficient proficiency region. Otherwise, he is said to have insufficient proficiency in the objective. Algebraically, if at the point of truncation, $w_m < (a_m + r_m)/2$, the examinee is said to possess sufficient proficiency in the objective; otherwise, he is said to lack adequate proficiency.

3.2 The Branching Component

In the last section, discussion centered upon classifying an examinee with respect to his proficiency in a specific objective. Assuming that the preceding plan represents a viable approach for making classification decisions regarding an examinee's proficiency in an objective, the next major concern of the test constructor is the development of a branching strategy. As noted before, the advantage of branching is that it permits the test builder to capitalize on his knowledge of the prerequisite relationships among a set of objectives in order to obtain an accurate profile of the examinee's competencies while testing as few of the objectives as possible.

A branching strategy was devised using the rationale that if two examinees have both evidenced proficiency in an objective, but one has made almost no errors, whereas the other has responded incorrectly to several items, the examinee who made fewer errors should be branched for testing on a more difficult objective than the examinee who made several mistakes. Likewise, it was believed

that the examinee who had insufficient proficiency in an objective and who answered nearly all of the items incorrectly should be branched for testing on an easier objective than the examinee who was classified in the same way but responded correctly to a larger proportion of the items that were presented to him. Table 3.1 summarizes the branching rules employed in the test model. Although the branching procedures are formally structured, the test builder should have freedom to adjust them so that conditions unique to a particular hierarchy can be accommodated.

Table 3.1

Branching Rules for Computer-Assisted Placement Testing

Decision for 1 Skill	Pupil's Response Data (p)	Branching Rules (Next Skill to be Tested)
Sufficient Proficiency ($p \geq .85$)	HIGH ($p \geq .93$)	Branch up to highest untested skill.
	LOW ($.85 \leq p \leq .93$)	Branch up to skill midway between this skill and highest untested skill
Insufficient Proficiency ($p \leq .60$)	HIGH ($.43 \leq p \leq .60$)	Branch down to skill midway between this skill and lowest untested skill.
	LOW ($p \leq .43$)	Branch down to lowest untested skill.

3.3 The Item Generation Component

Under the assumptions of the test model, the number of items required for testing an individual on a given objective could range from one, to the arbitrary value chosen for test truncation. The test items must be selected by randomly sampling from among the population of items that define the objective.

Both of the preceding observations suggest that it would be advantageous to store item generation code in the computer, thus permitting random generation of any item from the population of items for the objective. Such a procedure would be preferred to one that requires storing the entire population of items on the computer and then randomly sampling from among them. For many objectives, the size of the population of items would make the latter strategy impossible. An algorithm that facilitates construction of a random sample of items from a specified population is called an item generator. There are many advantages to using item generators. They do not require the use of huge amounts of computer memory, nor do they artificially restrict the size of the item pool that can be accessed by the test builder. Equally important, if the tests are to be used in an individualized instructional program, unique but equivalent forms of the tests can be generated in almost unlimited quantity (Ferguson and Hsu, 1971).

3.4 Rationale for Computer Implementation of the Test Model

It may be argued that many of the components of the test model previously described could be implemented without the direct aid of a computer. Although this may be a valid observation, it

is equally true that without computer assistance, the management problems associated with the administration of such tests would be overwhelming.

The ease with which a computer can randomly generate items according to user specification, integrate previously obtained test information with new data to make proficiency decisions subject to variable criteria, and then branch an examinee to test an objective appropriate for one having his competencies, is a strong argument for its use in implementing the test model. It has been suggested that the greatest potential use of the test model may be in individualized education settings where instructional decisions are made on the basis of test information. In such a situation, tests that are tailored to each examinee so as to produce a maximum amount of reliable information with a minimum investment of the student's and teacher's time, must function in the absence of elaborate schemes for their administration. This study proposes to demonstrate that a computer can provide the means for making a relatively complex test model tractable in just such a setting.

IV. Implementation of the Test Model

4.0 The Implementation Plan

As stated earlier (Section 1.2), the plan for this study was to, having completed the design of a computer-assisted branched test model for criterion-referenced measurement, examine its potential for increasing the effectiveness of pretesting and posttesting in IPI mathematics. Thus, the purposes of this chapter are to describe: (1) how test construction proceeded, (2) how the test was implemented, and (3) the nature of the data collected.

4.1 Development of a Test Hierarchy

Level D Addition-Subtraction was chosen as the unit to which the test model would be applied. Since the model requires the existence of a hierarchy defining the prerequisite relationships among the unit objectives, such a structure was hypothesized. Because a valid hierarchy is essential if accurate test profiles are to be generated, the proposed structure was examined intensively. A pilot study was undertaken to test its validity and provide information that could, if necessary, assist in its restructuring. Fifty-six IPI students were each administered two forms of a paper and pencil test on the unit objectives, thus generating sufficient data to restructure the hierarchy. The resulting unit was comprised of 18 objectives. Table 4.1 presents a verbal statement of each of the objectives; whereas, Figure 4.1 provides a graphic representation of the hierarchy for the same objectives. The number in each square

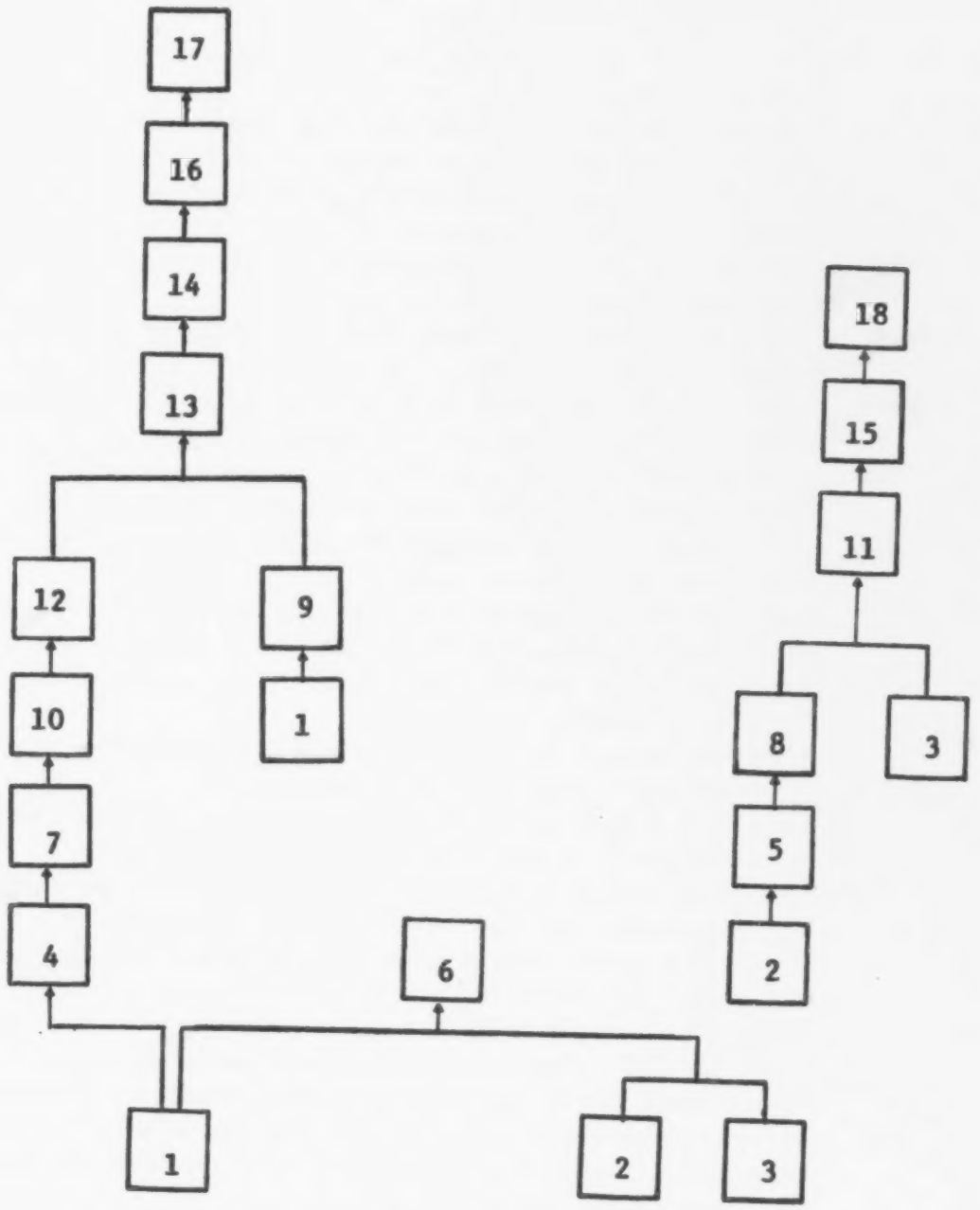
Table 4.1

Objectives for Level D Addition-Subtraction Unit

OBJECTIVE	
1	Solves addition problems from memory for sums less that or equal to twenty.
2	Solves subtraction problems from memory for sums less than or equal to nine.
3	Solves subtraction problems from memory for two digit sums less than or equal to twenty.
4	Solves addition problems related to single digit combinations by multiples of ten.
5	Solves subtraction problems related to single digit combinations by multiples of ten.
6	Finds the missing addend for problems with three single digit addends.
7	Does column addition with no carrying. Two addends with three and four digit combinations.
8	Solves subtraction problems with no borrowing. Three and four digit combinations.
9	Finds the sum for column addition using three to five single digit addends.
10	Does column addition with no carrying. Three or four digit numbers with three to five addends.
11	Subtracts two digit numbers with borrowing from the tens' place.
12	Adds two digit numbers with carrying to the tens' or hundreds' place. Two addends.
13	Adds two digit numbers with carrying to the tens' or hundreds' place. Three or four addends.
14	Adds two digit numbers with carrying to the tens' and hundreds' place. Two to four addends.
15	Subtracts three digit numbers with borrowing from the tens' or hundreds' place.
16	Adds three digit numbers with carrying to the tens' or hundreds' place. Two to four addends.
17	Adds three digit numbers with carrying to the tens' and hundreds' place. Two to four addends.
18	Subtracts three digit numbers with borrowing from the tens' and hundreds' place.

Figure 4.1

Hierarchy of Skills for the Level D
Addition-Subtraction Unit



of Figure 4.1 can be placed in one-to-one correspondence with the objectives listed in Table 4.1. The validity of the structure underwent examination after the computer assisted test designed for the study had been administered to a select sample of students. Those results will be reported later.

4.2 Development of Item Generation Rules

Construction of test items for a specific objective was accomplished by applying an item generation rule that was designed to randomly produce items from among the population of items for that objective. Figure 4.2 permits inspection of the algorithm used to construct items for objective 11, the latter requiring the subtraction of two digit numbers with borrowing from the tens' place. In the flow-chart, RANDOM is a function that generates a single random integer. Similar rules control the construction of items for each of the remaining 17 objectives.

It should be noted that stratification was involved in the generation of items for some of the objectives. For example, objective 7 requires the addition of two addends, each with either three or four digits. The generation of items for that objective was stratified so that items with three digit addends and items with four digit addends would be tested.

4.3 The Plan for Proficiency Classification

Values for p_0 and p_1, the criteria for sufficient or insufficient proficiency in a skill (described in Section 3.1), were

Figure 4.2

Flowchart of Item-Generation Rules for Objective Eleven

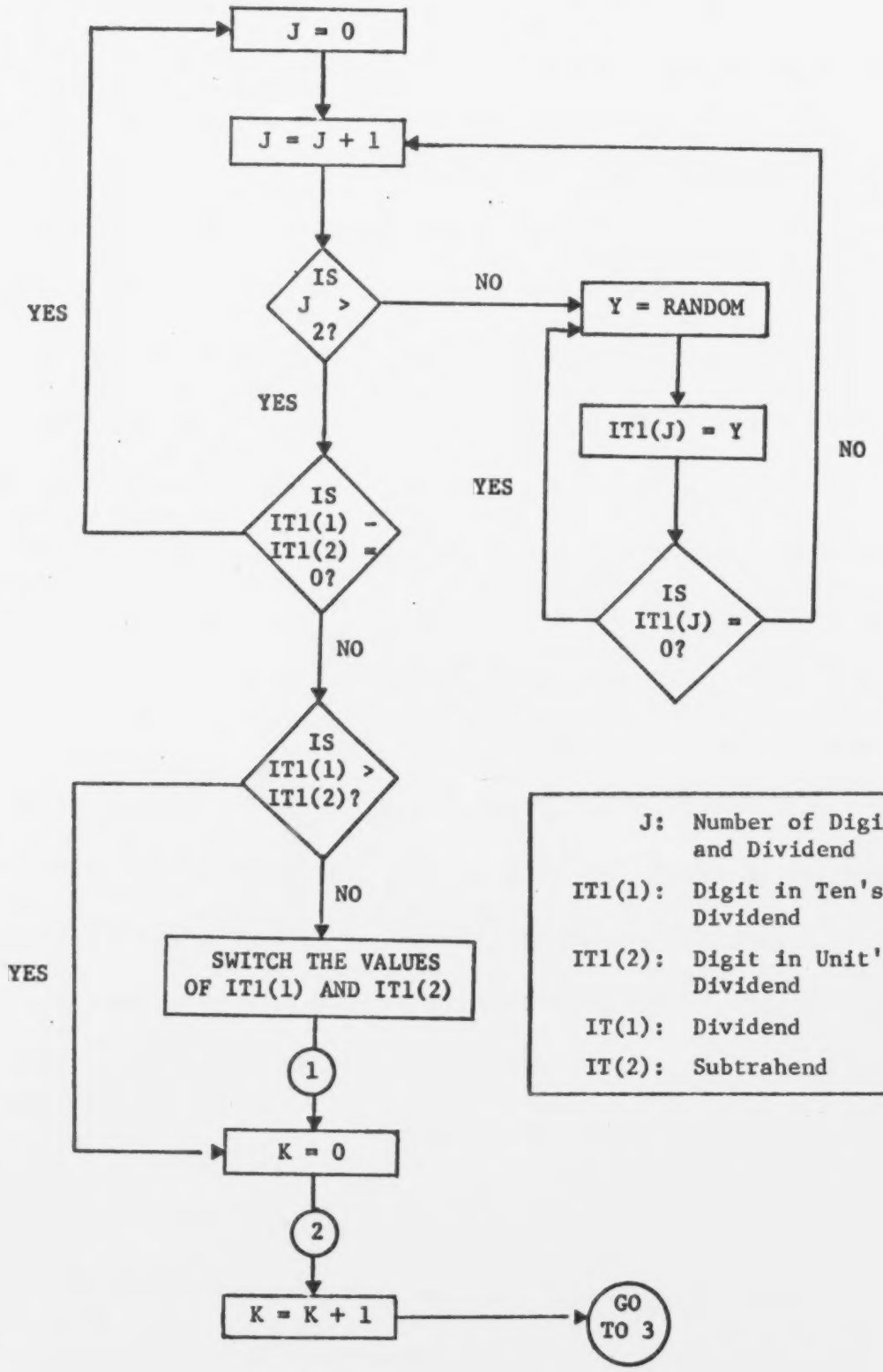
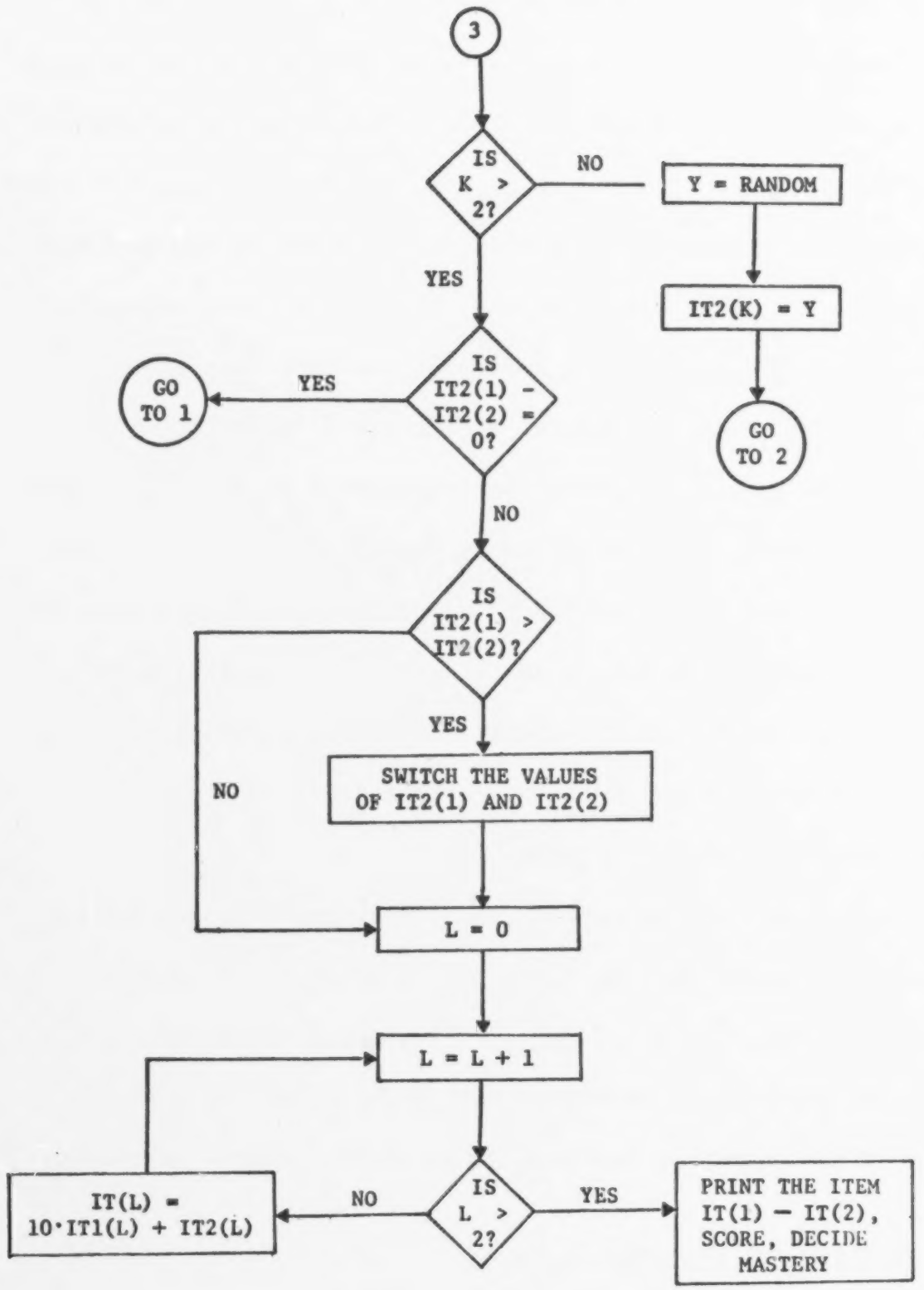


Figure 4.2 - Continued



selected after consultation with curriculum experts. Since several of the 18 objectives were judged to require a more strict set of criteria than others, two sets of values for p_0 and p_1 were used, one for each of two groups of objectives. For objectives 1, 2, 3, and 6, p_0 was set at .90 and p_1 at .70. The remaining objectives were tested with $p_0 = .85$ and $p_1 = .60$. The reason associated with the decision to place more demanding restrictions on the proficiency criteria for the first set of objectives was that they represented basic addition and subtraction facts, and as such were used extensively throughout the curriculum. Objective 6 was included in the first group because it involved the integration of two operations.

The values of α and β chosen for the study were .20 and .10, respectively. Since a Type II error was potentially more serious than a Type I error, β was chosen so that the risk of incorrectly deciding that the examinee had sufficient proficiency in an objective was substantially less than the risk of incorrectly deciding insufficient proficiency.

Note that for the second group of objectives, the values chosen for the study were $p_0 = .85$, $p_1 = .60$, $\alpha = .20$, and $\beta = .10$. For these values, Figure 3.2 (page 30) provides the means by which one can see the proficiency decisions that would result if m of the n items to which responses had been given, were answered correctly.

4.4 The Branching Strategy

Close inspection of Figure 4.1 (page 38) reveals that, for the unit at hand, the hierarchy on which testing should be based is comprised

of seven sequences. The latter are listed in Table 4.2. Each sequence consists of a set of objectives that are arranged so that beginning at the left, each objective is the prerequisite of all objectives to its right.

Table 4.2

Sequences for the Level D Addition-Subtraction Hierarchy

Sequence	Objectives Comprising the Sequence
1	1, 4, 7, 10, 12, 13, 14, 16, 17
2	1, 9, 13, 14, 16, 17
3	1, 6
4	2, 6
5	3, 6
6	2, 5, 8, 11, 15, 18
7	3, 11, 15, 18

Testing for all examinees began with objective 12 of the first sequence. Using the criteria described in Table 3.1 (page 33), the examinee was branched from objective to objective and tested until a classification decision was reached for each objective. For example, if it was determined that the examinee had sufficient proficiency in objective 12, as indicated by correct responses to 98% of the items presented, he was branched for testing on objective 17. If he had sufficient proficiency but responded correctly to less than 93% of the items presented, he was branched instead to objective 14 where testing continued. Either of the preceding decisions

classifying an examinee as sufficiently proficient in objective 12, coupled with a knowledge of the unit structure, would eliminate the need for testing objectives 1, 4, 7, and 10.

Exceptions to the branching strategy outlined in Table 3.1 occurred only when some justification for the action existed. For example, whenever a student failed to attain sufficient proficiency in objective 12 and also responded correctly to fewer than 43% of the items presented, he was branched for testing on objective 4 rather than objective 1, since previous testing experience had indicated that nearly every student who takes the unit test demonstrates sufficient proficiency in objective 1. Thus, objective 1 would be tested only if the examinee did not possess sufficient proficiency in all other objectives in the sequence.

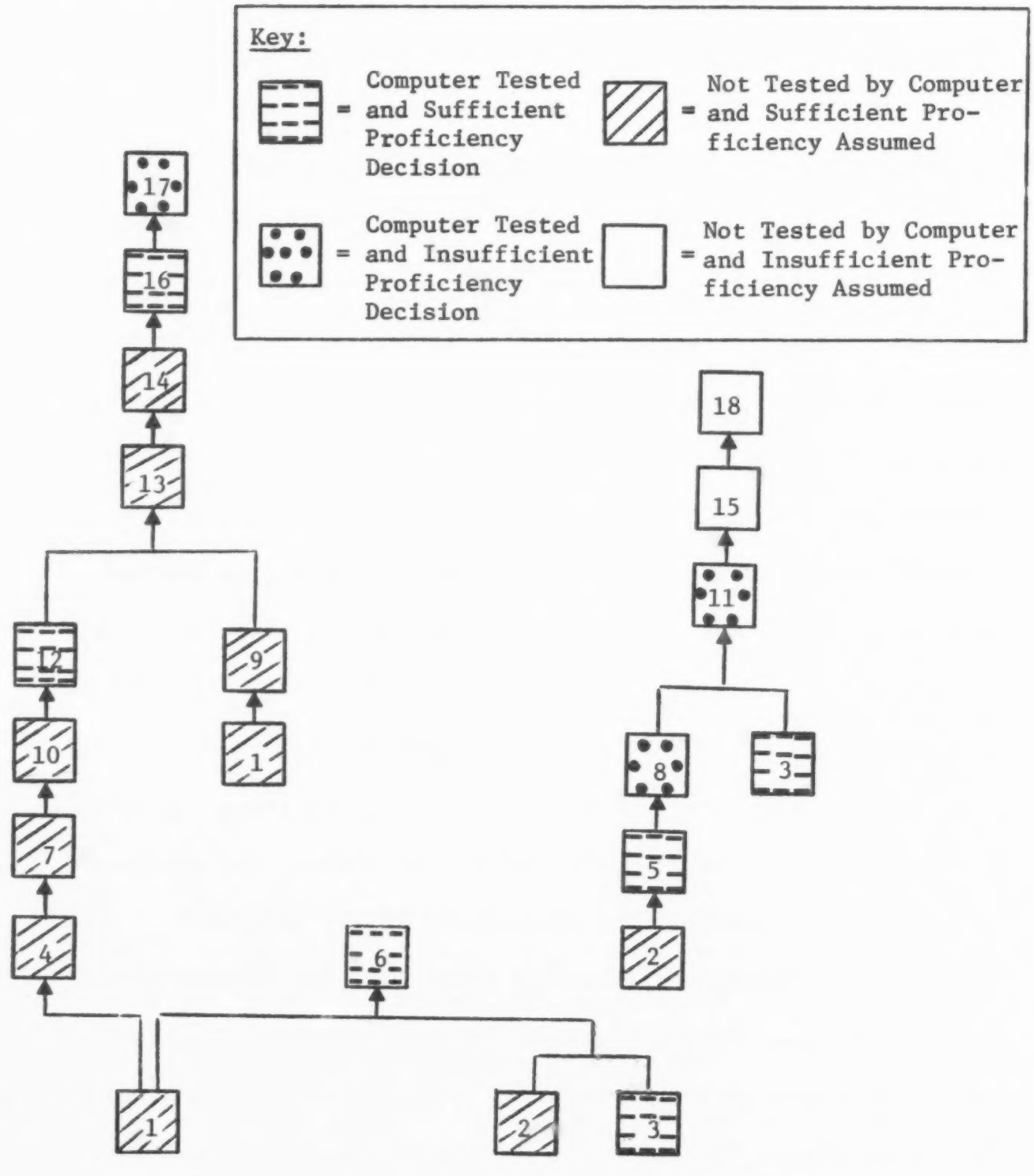
A sample of a profile that resulted from using the branching procedure employed by the test model is displayed in Figure 4.3. For that profile, note that the examinee was actually tested only on objectives 3, 5, 6, 8, 11, 12, 16, and 17. Thus, of the 18 objectives in the unit, only 8 were tested. Even so, classification decisions related to the examinee's proficiency in the remaining 10 objectives were possible.

4.5 Administration of the Test

Seated at a teletypewriter, the examinee provided a seed for a random number generator by typing his student ID and the date. He was then administered a single item that was randomly generated

Figure 4.3

Example of a Student Profile Resulting from the Computer-Assisted Branched Test



from among the population of items for objective 12. After he responded to the item, the computer immediately scored his response and then executed the routine that was programmed to determine whether or not he had demonstrated sufficient proficiency in the objective according to the specified criteria. As discussed before, if the examinee answered the first two items incorrectly, H_0 $(p=p_0)$ was rejected and H_1 $(p=p_1)$ was retained. In this case, he was said to have insufficient proficiency in the objective. The shortest route leading to a decision confirming sufficient proficiency when $p_0 = .85$ required a correct response to each of the first six items. In general, testing continued with the generation and presentation of another item whenever $.60 < p < .85$. Since it was possible for an examinee's true proficiency to lie in the "no-decision" region, a rule was applied for truncating the testing cycle after 30 items. In the event of truncation, the objective was said to be mastered if, after the thirtieth item, the decision function could be described as having fixed the point (m,w) on the graph (Figure 3.2) closer to the sufficient proficiency region than to the insufficient proficiency region. Otherwise, the examinee was said to have insufficient proficiency in the objective.

Once a decision was reached concerning the examinee's proficiency in a particular objective, he was branched for testing on another objective in the hierarhcy. The sequence of objectives on which a particular student was tested was a function of his unique proficiencies, the unit hierarchy, and the branching rules described

in Table 3.1. Testing was terminated when classification decisions were made for all objectives. A summary of the examinee's test performance was then generated to assist in formulating instructional strategy.

4.6 The Collection of Data

During the Spring of the 1968-1969 school year, the level D Addition-Subtraction test was administered to a sample of 75 students in grades one through six at the Oakleaf Elementary School. On two separate occasions each student was given the computer test at a teletypewriter. In most cases, each examinee took the two tests on consecutive days. In no instance did an examinee have instruction on the unit between tests. Since items were constructed using item generators stored in the computer, each test was unique, and the subtests on each objective were equivalent across tests.

Because there was likely to be a marked variation in the branching routes and test characteristics for individuals at the extremes of the unit proficiency continuum, three groups were identified for testing. The IPI curriculum coordinator identified 10 students for whom the probability of testing out of the unit was very high, another 10 for whom the probability of sufficient proficiency in any of the unit objectives was nearly zero, and 55 for whom performance expectations were between these two extremes. Of the 75 students tested, 25 had not yet entered the unit, 11 were currently working in the unit, and 36 had completed the unit at an

earlier date. Of the 28 who had not entered the unit, 10 were members of the group for whom expectations were low. Likewise, of the 36 who had completed the unit, 10 were members of the high proficiency group.

The field test was undertaken to permit the formal collection of data that could be used to address the following concerns:

(1) What are the implications for test length (number of items) and time to completion when an item sampling procedure that calls for controlling classification errors is employed?

(2) What does the branching procedure imply about test length (number of items) and time to completion?

(3) How do the item sampling strategy and the branching procedures employed by the test model affect test reliability?

Additional data were collected for the purpose of validating the hierarchy on which the branching was based.

Since the tests were administered by teletypewriter, a complete record of each test was preserved. As the examinee worked at the teletypewriter, a record of each objective tested was maintained. After completing the computer test, he was required to take a paper and pencil test on all objectives not directly tested by the branched test. The decision parameters used on the paper and pencil tests were the same as those used on the computer test. Thus, a measure of the examinee's proficiency in <u>all</u> of the unit objectives was recorded.

V. Evaluation of the Test Model

5.0 Validity of the Hierarchy

Since all examinees were tested on each of the 18 objectives, either by computer or with a paper and pencil simulation of the computer test, the resulting test profiles were used to ascertain the validity of the hierarchy that had been used to make branching decisions. Examination of the test profiles revealed very few inconsistencies. The latter occurs when, given two objectives A and B, with A prerequisite to B, an examinee is classified as having sufficient proficiency in B while having insufficient proficiency in A. A validity index was computed as the ratio between the observed number of inconsistencies and the total number that could have occurred. The value of the index, computed over 110 profiles, was .002. That is, of the total number of inconsistencies possible, only .2 percent actually occurred. With adequate support for the validity of the hierarchy for the sample at hand, meaningful consideration could be given to the test results.

Note that data for the two groups of size 10 were not included in the computation of the validity index, nor will they be included in most of the analyses to follow. This action was taken since all 10 children in the low proficiency group failed to achieve sufficient proficiency in all objectives. Similarly, all 10 children in the high proficiency group attained sufficient proficiency in all objectives. The use of this data would have a spurious effect on any indices reported. Further, the middle

proficiency group more closely reflects the ability of the potential user population for the test.

5.1 Validity of the Test

Whether or not the test measured what was intended that it should measure was an issue that was not difficult to resolve since the objectives that were tested were defined in precise behavioral terms. Moreover, the procedure used to construct the test items assured the existence of high content validity.

In another sense, the test could be valid only if it reflected an accurate measure of the examinee's proficiencies. Since inferences were made about objectives that were not tested, an important concern was the accuracy with which the branched test predicted an examinee's performance on those objectives. By matching the classification decisions reached for all objectives not tested by computer against the classification decisions that were reached using the results of the paper and pencil tests, it was possible to obtain an index of the extent to which the computer test possessed predictive validity in the sense described above. The proportion of correct proficiency classifications generated by the computer test was computed for each examinee. Then the mean of the entire sample of these proportions was computed. Since each examinee took the computer test twice, indices are reported for both test administrations. For samples of size 55, the two indices were .988 and .990, respectively. That is, the classification decisions reached by inference using the hierarchy to guide routing from one

objective to another, were found to be consistent with subsequent paper and pencil test outcomes approximately 99 percent of the time.

5.2 Reliability of the Test

Assuming that the test and structure were valid, it remained to be demonstrated that the test was also reliable. It can be argued that the test instrument was reliable only to the extent that the two profiles obtained for each examinee were in one-to-one correspondence. If, for all examinees, the two profiles were identical, one could infer the existence of a perfect relationship between test and retest classification decisions. Should there be gross discrepancies between profiles, one would be justified in expressing reservations about the reliability of the classification decisions. A necessary condition for this approach was that no instruction involving the unit occur between administrations of the test.

One procedure used to determine an index for placement reliability was to, on the basis of the computer test profiles, assign a score to each student on each of the seven linear sequences of skills comprising the unit. The latter were previously reported in Table 4.2 (page 43). For example, if the examinee had sufficient proficiency in objectives 1, 4, 7, and 10 of sequence 1 and insufficient proficiency in objectives 12, 14, 16, and 17, he was assigned a score of four for sequence 1. Once this process was repeated for all seven sequences for each examinee, the scores for the first test of

each sequence were correlated with the scores on the corresponding sequence obtained from the second testing. The resulting correlation coefficients are reported in Table 5.1. They indicate that there was a high relationship between classification decisions for each objective from one test administration to the next.

Table 5.1

Correlation Coefficients Between Repeated Measures of Proficiency for the Seven Linear Sequences (N=55)

Sequence Number	r
1	.95
2	.90
3	.83
4	.81
5	.91
6	.96
7	.96

From an instructional point of view, it is probably more meaningful to examine test reliability from still another frame of reference. Since it is assumed that an examinee's placement in each sequence determines his instruction, it is of interest to know the proportion of the 55 students who would receive the same instruction after completing the computer test twice. Table 5.2 reports the proportion of the 55 examinees for whom instruction would have differed by 0, 1, 2, or 3 objectives. In no case did two profiles differ on more than 3 of the objectives. Nearly 80

percent of the examinee's would receive instruction with one or fewer differences required as a consequence of the retest.

Table 5.2

Proportion of Students for Whom Instruction Would Vary from Test to Retest (N=55)

Number of Variations in Objectives from Test to Retest	p
0	.45
1	.34
2	.20
3	.01

Table 5.3 provides essentially the same information as Table 5.2 but reports it by sequence. As can be observed, it was rare that retesting yielded a difference of more than one objective on any sequence.

Table 5.3

Proportion of Students Differing from Test to Retest in the Number of Objectives for which Instruction on a Sequence was Required

Sequence Number	Difference in Number of Objectives 0	1	2	3
1	.60	.29	.09	.02
2	.58	.34	.05	.03
3	.93	.07	.00	.00
4	.93	.07	.00	.00
5	.93	.05	.02	.00
6	.78	.22	.00	.00
7	.84	.16	.00	.00

A recurring problem seen within the conventional IPI testing program is the frequent inconsistency in measure from pretest to posttest. It often happens that an examinee is found to have sufficient proficiency in an objective on a pretest only to be recorded later as having insufficient proficiency on the first or subsequent posttests. The left half of Table 5.4 is a summary of such occurrences for a sample of IPI mathematics students in several schools. The N's vary since the number of examinees passing through the objective on the pretest also varies.

Table 5.4

Comparison of the Conventional Test with the Computer Test on the Consistency of Proficiency Classification From Pretest to Posttest

Skill	Percent of regression from pretest to posttest on the Conventional Test	N	Percent of Regression from Pretest to Posttest on the Computer Test	N
1	1	431	0	53
2-3	5	539	0	53
5	13	333	2	53
6	12	390	0	53
7	13	491	2	53
8	26	251	2	53
10	14	201	2	53
12	17	145	0	53
14	23	156	0	53
15	27	154	4	53
16	67	12	0	53
17	36	56	6	53

The right side of the table presents a similar summary for computer-assisted test data. The skills for the conventional tests have been matched on a one-to-one basis with corresponding skills on the computer test and both appear at the same level in the table. Recall that the unit used for the computer assisted test included objectives not in the original unit (Section 4.1). Although the number of days between pretest and posttest for the conventional test was several days greater than for the computer test, in neither instance did any formal instruction on the skill take place between tests. The N was fifty-three for the computer-assisted test since any of the 55 students who would have pretested out of the unit were not included in the analysis.

With the time delay between tests a noted restriction, one can at least say that the results offer encouragement that by employing an item sampling technique that permits control over classification errors, the computer-assisted test model may increase test reliability.

5.3 Test Length and Time-To-Completion as a Function of Item Sampling

One reason for developing the computer-assisted test model was the belief that, when employed, it could substantially reduce the amount of time required to obtain accurate information about a student's proficiencies in a set of linearly related skills. Of course, the time-to-completion of a test is directly related to the number of items presented during the course of the test. Since the number of items presented to a particular examinee on the computer

test was a function of both the item sampling technique and the branching procedures employed, it would seem wise to separate the two effects. Discussion will first be focused upon the implications of item sampling for test length; that is, the total number of items presented per test.

Table 5.5 provides data for comparing the mean number of items presented to the 55 examinees who took the computer test with other examinees who had been administered the conventional test.

Table 5.5

Comparison Between Conventional and Computer Tests on the Number of Items Presented (N=55)

Objective	Number of Items Presented on the Conventional Test	Mean Number of Items Presented on the Computer Test			
		Test I		Test II	
		$\bar{X}$	s	$\bar{X}$	s
1	40	9.8	3.45	9.6	1.97
6	5	10.5	6.03	9.3	4.55
7	6	6.6	2.48	6.4	2.52
8	5	7.2	4.31	6.1	2.44
9	6	7.0	2.27	7.8	3.76
11	5	4.9	4.02	4.7	4.36
14	6	7.2	4.97	7.5	6.22
15	5	4.6	4.58	4.8	4.77
16	4	7.0	5.44	8.3	6.83
17	4	6.8	6.39	9.4	8.29
18	5	5.3	6.42	5.2	5.56

In order that data be available for all 18 objectives for all examinees who took the computer test, and since many of the objectives were not tested by computer due to the branching procedure employed by the test, recall that examinees were administered paper and pencil tests on all objectives for which they had not been tested by computer. The latter tests employed the same item sampling and classification procedures used by the computer test model. The table includes only those objectives for which comparison was possible given the data available. The data in the table reveal that, on the average, the item sampling procedure employed by the computer test required that more items per objective be tested than did the conventional tests.

It is also of interest to compare the distribution of the number of items required to reach a proficiency decision for a given objective on the computer test with the number of items required by the conventional test. Table 5.6 reports the proportions of examinees for whom a smaller, equal, or larger number of items was required to reach a classification decision on the computer test than for the comparable objective on the conventional fixed length test. Of the 11 objectives that could be compared, the data indicates that only 3 required fewer items when administered by computer as opposed to conventional paper and pencil tests. Simultaneous comparison of Table 5.5 with Table 5.6 yields some insight into the outcomes evident in Table 5.6. Note that the conventional test for objective 1 required 40 items, a considerably more stringent

requirement than placed upon the same objective by the criteria selected for use on the computer test. Objectives 11 and 15 proved to be very difficult for the examinees. Consequently, many examinees failed to evidence proficiency in the objective and computer testing terminated quickly.

Table 5.6

Proportions of Examinees Receiving A Smaller, Equal, or Larger Total Number of Items on the Computer Test than on the Conventional Test (N=55)

Objective	Proportion of Examinees		
	Fewer	Equal	More
1	1.00	.00	.00
6	.16	.00	.84
7	.10	.00	.90
8	.16	.00	.84
9	.05	.67	.28
11	.53	.02	.45
14	.31	.37	.32
15	.60	.00	.40
16	.30	.00	.70
17	.40	.00	.60
18	.04	.04	.32

To summarize, the data presented thus far in this section indicate that the computer test did, on the average, require that more items be tested per objective than did the conventional fixed length test. However, it must also be noted that the criteria used

for classification decisions on the computer test (see Section 4.3) guard against classification errors more stringently than do the conventional tests. Further, the number of items comprising the paper and pencil tests were determined somewhat arbitrarily, making this comparison of interest only to those who use the conventional IPI tests.

A brief summary of the data available for the two small groups comprised of children with extreme proficiencies is also of interest. For the group with 10 examinees who had minimal proficiencies in the objective, the computer test resulted in the termination of testing on an objective long before the conventional fixed length test. Using the criteria specified in Section 4.3, if an examinee lacked sufficient proficiency in an objective, the computer might terminate testing in the objective after 2 items; whereas, the conventional test usually required a response of 4 or 5 items.

For the small group of students who were highly competent in the objectives, the computer test compared favorably with the conventional test in terms of the number of items required during testing. For most of the objectives, the former required that a minimum of 6 items be presented to ascertain sufficient proficiency. Consequently, for this group, the average number of items presented per skill was only one or two larger than that required for the conventional test.

5.4 Test Length as a Function of Branching

Operationally, when a child takes a pretest or posttest in IPI mathematics, he is tested on all objectives comprising the unit. Consequently, for the unit at hand, 18 objectives would require testing. By taking advantage of a knowledge of the unit hierarchy, it is possible to substantially reduce the number of objectives that are tested. Table 5.7 provides the mean number of objectives on which the three groups, each of varying proficiency, were tested.

Table 5.7

Mean Number of Objectives Tested by Computer for Groups of Varying Proficiency

Group Proficiency	Test I			Test II		
	$\bar{X}$	s	N	$\bar{X}$	s	N
Low	7.00	0.00	10	7.00	0.00	10
Middle	7.51	2.31	55	7.35	2.25	55
High	5.00	0.00	10	5.00	0.00	10

The branching design fixed the lower and upper bounds for the number of objectives that could be tested at five and ten, respectively. Thus, the examinee whose profile was the most difficult to complete required testing on ten objectives, 55% of the number required by the conventional instrument. The individual who demonstrated unit mastery might be tested on as few as five objectives; that

is, only 28% of the number required by the conventional test. Fifty different branching routes were followed during the course of the study, thus clearly establishing the flexibility of the test model to adapt to individual differences.

It should be noted that examinees who had sufficient proficiency in all skills were tested on a minimum of seven objectives. By way of contrast, examinees who had sufficient proficiency in all skills were tested on a minimum of five objectives. This explains the means for the two extreme proficiency groups reported in Table 5.7.

Since the conventional tests now used in IPI mathematics do not incorporate the notion of skill hierarchies as an integral part of the testing process, the preceding comparisons may inflate one's impression of the success of the branching procedure in terms of its potential for minimizing testing time.

The computer is able to manage the branching process described in section 4.4 with far greater ease than it could be accomplished manually. However, it would be easy to adopt a manual branching strategy that calls for the examinee to start at the objective that was lowest in the hierarchy, say objective 1, and work his way up the structure until he arrived at an objective in which he was unable to demonstrate sufficient proficiency. For an example, refer to Figure 4.1 (page 38). If the examinee was tested on objective 1 and found to have sufficient proficiency, he would be branched to objectives 4 and 6 for testing. As long as he continued to demonstrate sufficient proficiency, he would be branched upward to another

objective, one step at a time. A similar procedure would be followed for the other sequences in the hierarchy; namely, those beginning with objectives 2 and 3.

Table 5.8 provides for a comparison of the efficiency of the branching procedure just described with the binary branch strategy used by the test model in the study. Data for the one step ladder entry in Table 5.8 was generated by applying that procedure to the existing profiles for the 55 students comprising the middle proficiency group. Whether or not the outcomes would have differed had the procedure been real, rather than simulated, is a matter for speculation. However, to the extent that previous testing was reliable, the data in Table 5.8 permit some interesting comparisons. It is dramatically clear that, for the objectives and sample of students at hand, the one-step ladder was quite inferior to the binary branch strategy. The latter method required that approximately 50% fewer objectives be tested than the former.

Table 5.8

Number of Objectives Tested Using Two Different Branching Strategies (N=55)

Branching Procedure	Test			
	I		II	
	$\bar{X}$	s	$\bar{X}$	s
Binary Branch	7.5	1.5	7.4	1.5
One Step Ladder	14.1	4.0	14.6	3.9

It seems reasonable to test the notion that a more optimal branching strategy might have been used in the study. Consequently, two other strategies were simulated using the profile from the previously administered tests. The purpose of the simulation was to ascertain whether either routing method would reduce the number of objectives to be tested and still yield the same unit profile. The branching rule for the first technique required that the examinee be branched up one objective in the sequence if he was found to be sufficiently proficient in the objective being tested and down two objectives if the reverse was true. The second technique was the converse of the first, branching up two objectives in the sequence if the examinee attained sufficient proficiency in the current objective, while branching down one objective in the other instance.

The outcomes of the simulations show that the procedure used for the tests in the study was markedly superior to either of these two methods. In 150 trials, the first strategy required the testing of fewer objectives only 11 times, the same number of objectives 47 times, and more objectives in 92 cases. The second procedure, although better than the first, was still not as efficient as the one used in the study. It required fewer objectives 36 times, the same number 24 times, and more objectives 90 times.

5.5 Some Observations About Implementation of the Computer Assisted Test Model

The test was administered using the IBM Model 360/50 computer with the University of Pittsburgh Time Sharing System. Although most

of the children had no previous experience with computer testing or instruction, they were able to take the tests on the teletype-writer with a minimum of direction and supervision. Since the test required that only the integer and control keys be used, other areas of the keyborad were covered. Provision was made for the examinee to alter a response if he had entered it incorrectly. Such corrections took place on the average of about one every two tests. System response time between entry of a solution to one item and the presentation of the next item was nearly always excellent. There was seldom an observable time lag.

The test was programmed so as to produce a printed summary of the student's proficiency in each objective in the unit. Additional item information could easily be added. An example of the entire output for a computer-assisted branched test is found in Appendix A. Teachers and supervisors were very enthusiastic about the potential use of such tests for pretesting and for diagnostic purposes. Particular interest was expressed in the use of such instruments for students with physical and learning disabilities. Several students fitting the latter descriptions were included in the test sample. In each case, the child was able to handle the testing without noticable difficulty and, in fact, seemed to enjoy working with the computer.

No serious attempt was made to evaluate the computer-assisted branched test in terms of cost effectiveness. However, it is clear that the computer would eliminate the need for a teacher or aide to

score the test and to control access to the test. Further, it provided the means to generate a nearly inexhaustible supply of unique but equivalent tests.

One of the major advantages of the computer test is the flexibility that it permits for varying the criteria for classification decisions; specifically, p_0, p_1, α, and β. The values of p_0 and p_1, although held constant from one test to the next in this study, were varied from objective to objective. α and β were fixed at .20 and .10, respectively. Since the number of items an examinee is given for an objective is a function of the preceding four parameters, the length of time required to complete a test is greatly affected by changing them. Thus, the choices for these values must be influenced by practical considerations. Some thoughts on how this can be accomplished are reported later in the report.

It should be noted that frequently an examinee's proficiency in an objective was greater than the criterion p_0 that was specified for sufficient proficiency. This was a common occurrence for students in the high proficiency group. Likewise, it was often the case that an examinee's proficiency in an objective was less than the criterion p_1 selected for insufficient proficiency. Since the item sampling technique employed in the study tested the hypothesis $p = p_0$ against the hypothesis $p = p_1$, the question arises as to whether or not the error rates α and β hold in the event of either of the preceding occurrences. However, α and β are upper bounds for error of Type I and Type II, respectively. Thus, in all cases, the error rates are applicable to the testing undertaken in the study.

It was noted earlier than if an examinee had responded to thirty items and no decision about proficiency had been made, the test was truncated, a classification decision made, and the examinee branched for testing on another objective. During the course of testing, it was necessary to truncate testing on an objective only six times. On four of these occasions, the decision involved a terminal objective. The risk of a misclassification has no greater implications for instruction in this situation than it does if the misclassification is made when the testing process terminates naturally. The seriousness of the error may even be smaller since in the event of a truncation, the routing strategy called for branching to an objective that was never more than one level of difficulty away from the objective being tested.

VI. Implications for Continuing Research in Models for Computer Assisted Testing

6.0 Overview

The preceding chapter of this report examined the reliability and the validity of a computer-assisted test developed for a unit in IPI mathematics. Further, it provided data that were useful in contrasting the automated version of the test with the conventional paper and pencil tests. In many respects, the computer test appeared to be comparable or superior to the latter. The branching procedure used by the computer test produced the most impressive effect, namely, a substantial reduction in the amount of time typically required to complete a unit pretest or posttest. Should a decrease in the amount of time required for testing prove to be a typical by-product of computer testing procedures, the implications for instruction are many. During the course of a school year, large numbers of hours now spent in testing, could be invested in instructional activities.

Assuming that replication of the study would yield approximately the same results for other units, a more intensive investigation of the computer-assisted test model is warranted. The remainder of this report is addressed to issues that were not resolved by the study and to some observations regarding how further development and refinement of the test model should proceed.

6.1 Suggested Refinements for the Test Model

Implementation and evaluation of the computer test have served to identify components of the test model that could be

substantially improved. They include adjustments in procedures for: (1) item selection and generation, (2) specification of the values for parameters that govern the classification rules used to determine the state of an examinee's proficiency, and (3) routing an examinee from one objective to another.

Section 4.2 of this report described the method by which items that tested a single objective were generated. The procedure called for the construction of items that constituted a random sample from the population of items that defined the objective. This was accomplished by using an item generation rule that was stored in the computer. The weakness of such a procedure is that, although it yields a random sampling of items from the population, it does not necessarily produce a representative sampling of those items.

If the content of a particular objective is analyzed, it soon becomes clear that many different forms of items may be required if testing of the objective is to be adequate. A grossly over simplified example will serve to amplify the problem. The problems 0 + 8, 2 + 8, and 9 + 8 are all representative of the behavior described by an objective that requires the sum of two single digit addends. Although all three items fit the objective, they represent different dimensions of the behavior described by that objective. It is conceivable that a random sampling of items from among the population of single digit combinations might yield no items of the first form, that is, items for which one addend is

zero. One way to resolve this problem is to expand the number of objectives so as to include each item form as a single objective. This is likely to be a mountainous and unrewarding task. If objectives were stated so that each could be tested using a single item form, this might imply that instruction should be similarly micro-oriented. That is, each instructional sequence would tend to focus on instruction for a single precise behavior, probably to the exclusion of how that objective was related to other objectives in the curriculum. A better approach is to analyze each objective for the purpose of identifying all of the item forms that one can recognize as belonging to the population defined by the objective and assure that items representative of each of these forms are included on the test. In this way, the items generated for a particular objective are both randomly constructed and representatively sampled.

From an instructional frame of reference, it is extremely important that testing be representative. Failure to adequately test an objective may result in a student's advancement to objectives requiring prerequisite objectives that he does not possess. If a test of the simple objective stated earlier, adding two single digit numbers, includes 20 items, one must be certain that items which include addition with zero as one of the addends are among the 20. Further, the computer should spotlight the specific form(s) of items that an examinee is unable to solve. Passing a student along to another objective because he responded correctly to 18

of 20 items testing the objective may not be an appropriate action even if the criteria for sufficient proficiency is .90. If it happens that the two items answered incorrectly involved the item for 0 + X where $X \varepsilon \{1,2,3,. . ., 9\}$, then regardless of the fact that the examinee satisfied the criteria for sufficient proficiency in the objective, instruction on how to solve problems representative of that item form should be required. The computer can serve as an effective agent for identifying the nature of the error and bringing it to the attention of the teacher.

The notions just described reflect the rationale behind a more refined approach for the selection and generation of items for computer testing than was used in this study. As a consequence of the work completed to date, a new procedure for domain referenced item generation has been adopted for use in future test development. The procedure calls for specification in behavioral terms of the objectives to be tested and analysis of each objective so as to identify item forms representative of all behaviors implied by each objective in the set. A single algorithm is developed that permits the construction of items representative of all item forms belonging to the set. The item generation technique employed for the 18-skill unit reported in this study required that 18 unique item generators be constructed. The improved procedures would reduce that requirement to a single, more comprehensive generator. A more detailed report of this strategy is reported by Ferguson and Hsu (1971).

One advantage of the item sampling technique employed by the computer test model is the flexibility it permits for varying the parameters p_0, p_1, α, and β. In the study, p_0 and p_1, the criteria employed for proficiency classification decisions were held constant for each objective among tests but were varied from objective to objective within tests. The values for α and β were fixed at .20 and .10, respectively.

Since $\underline{n}$, the number of items to which an examinee responded while being tested on a given objective, was a function of the preceding parameters, it would be possible to alter $\underline{n}$ by simply changing their values. The choices for these values must be influenced by practical considerations such as the expected number of items required in order to reach a classification decision for a given objective.

Specifying values for α and β can be resolved by embracing one of three alternatives: (1) minimizing the probability of a Type II error (assumed to be the more serious error) by reducing β, (2) reducing the number of items sampled for a given objective (at the cost of increasing the chance of errors of classification), or (3) selecting a middle role somewhere between the strategies presented by (1) and (2).

Although the first option could be taken, any advantage that the test model offers with respect to effectively reducing testing rates would be lost. At the other extreme, increasing the tolerance for classification errors by reducing $\underline{n}$ is likely to

yield information that is of little value for instructional decision making. Steering a course somewhere between these two extremes would appear to be the most viable approach. Examination of data like that found in Table 6.1 could prove useful in helping to redefine realistic but adequate values for protecting against Type I and Type II errors. Given the values for the parameters used in the study, Table 6.1 provides a summary of the average number of items required to test each objective. The data are based on two administrations of the test to each of the three proficiency groups. Since the item sampling and classification procedures were applied both to the objectives tested by computer and to those that were later tested using a paper and pencil format, both data are incorporated in the table.

Judging from the data reported, it would be possible to substantially reduce β for many of the objectives without increasing to intolerable levels the average number of items tested. For the middle proficiency group, objectives 4, 11, 15, and 18 appear, on the average, to require fewer items in order to reach classification decisions. Consequently, lowering β for these objectives is unlikely to increase these averages to unacceptable levels.

Although the study demonstrated that the branching strategy employed by the test model was extremely effective in eliminating the need for testing objectives for which classification decisions could be reached on the basis of hierarchical structure of the

Table 6.1

Average Number of Items Tested Per Objective
on the Computer Test

Objective	Group Proficiency		
	Low	Middle	High
1	4.6	9.7	9.0
2	4.8	10.2	9.0
3	2.0	9.6	9.0
4	2.4	6.1	6.0
5	2.0	6.6	6.0
6	2.0	9.9	9.0
7	2.0	6.5	6.6
8	2.0	6.7	6.0
9	2.2	7.4	6.2
10	2.0	7.4	6.6
11	2.0	4.8	6.0
12	2.0	6.7	6.0
13	2.0	6.5	6.4
14	2.0	7.4	7.0
15	2.0	4.7	6.4
16	2.0	7.6	7.0
17	2.0	8.1	8.2
18	2.0	5.2	6.2

unit, it is possible that the branching strategy might be made even more efficient. Consideration should be given to a branching approach that does not require that testing for all students begin with the same objective. The approach used in the study was to begin testing every examinee on objectives found in the middle

of the major sequences of the structure. Thus, every examinee was tested on objectives 12 and 13. Examinees who might easily have begun at some objective of greater difficulty were thus forced to solve problems far below their level of competency. Examinees who were incapable of solving problems at the level at which testing began, were forced to attempt to do so.

A possible solution to this problem is to permit the examinee to determine where testing of a sequence should begin. He would make such a decision by studying a sample item for each objective to be tested. If a Cathode Ray Tube (CRT) device were used, a sample of these items could be flashed on the screen upon request. After making a judgment of his own proficiencies in the given skills, testing would commence at the objective elected by the examinee. Such an option would be presented to him each time he was to enter a new sequence of objectives. Thus, for the D Addition-Subtraction unit reported in the study, he would be given two options, one for the addition sequence and one for the subtraction sequence. Investigation of such a branching plan would require a study of the accuracy with which examinees are able to identify the objectives in which they have competency.

6.2 Summary

The typical expression of reservation regarding branched testing has been with respect to its characteristic inability to improve upon conventional test measurement for the examinee of average ability. Although this sentiment has usually been expressed

with reference to normative measure, the extent to which it is equally true for criterion referenced measurement has been in doubt. The results of this study strongly suggest that this was not characteristic of the model tested in the study. To the contrary, given the objectives that were tested, it was extremely effective for the group with middle proficiency.

To a large extent, this can be attributed to the specific unit of work that was tested. For units with fewer objectives and with smaller scales, the effect of branched testing for the majority of examinees may be less pronounced. Nevertheless, this study has shown that computer-assisted testing can be used effectively for all students in an individualized instruction setting. Further, the measures yielded by such a procedure can be as valid and reliable as those for a conventional test with the additional bonus that they are obtained in less time with testing of fewer total items and objectives.

The item sampling procedures used by the computer test model affords the test constructor a much greater capability for controlling sampling error than is the case for conventional fixed length tests. Thus, improvement in the automated test is twofold. By tailoring the test to individuals, fewer objectives need to be tested and the objectives that are tested are less subject to errors of proficiency classification.

The item-sampling procedure used by the test model for determining an examinee's proficiency in an objective required

that the conditions for a Bernoulli experiment be met. Although each of the conditions was assumed to hold in the context of the study, it is recommended that an effort be made to test the assumptions of the model as they apply to specific objectives within a mathematics unit. Such a study is currently underway and a report of the outcome will soon be available.

Although the application of the test model to the pretest function in IPI mathematics proved highly successful, such a test does not provide adequate diagnostic information to suggest detailed instructional treatment for objectives in which a student is not competent. To the contrary, the objective of the branched test was to determine which objectives an examinee did not possess and to do it with as little testing as possible. Thus, it was desirable that he not be tested on objectives in which he lacked proficiency if such a judgment could be made on the basis of his prior performance and the hierarchy of prerequisite relationships among the objectives. Obviously, if an examinee was found to lack competency in an objective on which he was not tested, no precise data were available on which to base a decision regarding the best possible instruction for him. Diagnostic computer tests using the same item sampling procedures, but specific to single objectives, will provide the means whereby the computer can suggest instruction that is appropriate for each individual examinee.

Tests that are developed using the item sampling procedures suggested in Section 6.1 offer an additional advantage over tests

developed using conventional techniques. They serve as an agent for generating and collecting data needed to determine how precisely objectives and/or item forms need to be specified. A procedure that makes possible the rapid generation of items that are representative of any of a tremendous number of item forms, encourages investigation of the relationships among those forms. Further, the procedures used to select and generate items, aids in the refinement of the curriculum and instructional materials associated with it because it demands a thorough examination of the relationships among the objectives, how they are taught, and what is tested.

With the computer's power to permit storage and retrieval of large amounts of data, it is clear that computer-assisted testing can be a substantial contributor to an adaptive educational environment that permits experimentation concerned with optimizing the integration of information collection, computer testing, and instruction. The integration of these components into a functional unit that can effectively provide for the individual needs of a large number of students, will provide a challenging exercise for researchers. Some of the contributions that computer-assisted testing can make in such an instructional model have been brought into sharper focus by this study.

APPENDIX A

SAMPLE COMPUTER ASSISTED BRANCHED TEST

(Entire printout is condensed to a single page.)

PLEASE TYPE YOUR STUDENT NUMBER
>1684

TYPE THE SCHOOL DATE AS A NUMBER
>180

Page 1

OBJECTIVE 12

36
+17
>53

71
+80
>151

29
+21
>50

90
+33
>123

22
+39
>61

55
+50
>105

OBJECTIVE 17

456
+375
>831

Page 2

319
166
+185
>670

286
253
145
+121
>805

278
+399
>677

189
477
+282
>948

100
378
137
+140
>755

OBJECTIVE 11

92
-72
>19

Page 3

81
-36
>45

85
-57
>28

93
-87
>6

70
-59
>11

30
-18
>12

OBJECTIVE 18

970
-571
>339

910
-348
>562

911
-145
>766

Page 4

921
-189
>732

216
-139
>77

905
-249
>656

OBJECTIVE 6

8
?
+8
22
>6

8
3
+?
19
>8

?
6
+1
8
>1

Page 5

2
?
+5
12
>5

2
1
+?
5
>2

?
6
+6
19
>7

6
?
+9
23
>8

4
1
+?
12
>7

?
7
+8
17
>2

SUMMARY OF BRANCHED TEST FOR UNIT D ADDITION-SUBTRACTION

ALL OBJECTIVES SUFFICIENTLY MASTERED

REFERENCES

Angloff, W. H. and Huddleston, E. M. "The Multi-Level Experiment. A Study of a Two-Level Test System for the College Board Scholastic Aptitude Test," Educational Testing Service Statistical Report, 58-21. Princeton, New Jersey: Educational Testing Service, 1958.

Bayroff, A. G. and Seeley, Leonard C. "An Exploratory Study of Branching Tests," United States Army Behavioral Science Research Laboratory Technical Research Note, 188. Washington, D.C.: Military Research Division, 1967.

Bolvin, John O. IPI - A Manual for the IPI Institute, Learning Research and Development Center, Pittsburgh, Pennsylvania: University of Pittsburgh, 1967.

Cleary, T. Anne; Linn, Robert; and Rock, Donald. "An Exploratory Study of Programmed Tests," Educational and Psychological Measurement. XXVIII, 1968.

Cooley, William and Glaser, Robert. "The Computer and Individualized Instruction," Science, CDXVI, October, 1969.

Ferguson, Richard. "The Development, Implementation and Evaluation of a Computer Assisted Branched Test for a Program of Individually Prescribed Instruction," Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Ferguson, Richard. "Computer Assistance for Individualizing Instruction," Computers and Automation, XIX, March, 1970.

Ferguson, Richard and Hsu, Tse-chi. "The Application of Item Generators for Individualizing Mathematics Testing and Instruction," Publication Series, Learning Research and Development Center, University of Pittsburgh, 1971.

Flanagan, John. "Functional Education for the Seventies," Phi Delta Kappan, September, 1967.

Glaser, Robert. "Adapting the Elementary School Curriculum to Individual Performance," Proceedings of the Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1968.

Glaser, Robert. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist. XVIII, 1963.

Green, Bert F., Jr. "Comments on Tailored Testing," Report to the Conference on Computer-Based Instruction, Learning, Testing, and Guidance. Austin, Texas, 1968.

Hanson, Duncan N. and Schwarz, Guenter. "An Investigation of Computer-Based Science Testing," Report submitted to the College Entrance Examination Board. Tallahassee, Florida: The Florida State University, 1968.

Kriewall, Thomas E. and Hirsch, Edward. "The Development and Interpretation of Criterion-Referenced Tests," Paper presented at the annual convention of the American Educational Research Association, Los Angeles, California, February (1969)

Lord, Frederick. "Some Test Theory for Tailored Testing," Report to the conference on Computer-Based Instruction, Learning, Testing and Guidance. Austin, Texas, 1968.

Patterson, J. J. "An Evaluation of the Sequential Method of Psychological Testing," Unpublished doctoral dissertation, Michigan State University, 1967.

Suppes, Patrick. "Modern Learning Theory and the Elementary School," American Educational Research Journal. I, 1964.

Wald, Abraham. Sequential Analysis. New York: John Wiley and Sons, Inc., 1966.

Waters, Carrie Jean. "Preliminary Evaluation of Simulated Branching Tests," United States Army Personnel Research Office Technical Research Note, 140. Washington, D.C., 1964.

DISTRIBUTION LIST